Nonparametric Evaluation of Noisy ICA Solutions

Syamantak Kumar[1] **Purnamrita Sarkar**[2] **Peter Bickel**[3] **Derek Bean**[4]

[1]Department of Computer Science, UT Austin
[2]Department of Statistics and Data Sciences, UT Austin
[3]Department of Statistics, University of California, Berkeley
[4] Department of Statistics, University of Wisconsin, Madison
syamantak@utexas.edu, purna.sarkar@austin.utexas.edu,
bickel@stat.berkeley.edu, derekb@stat.wisc.edu

Abstract

Independent Component Analysis (ICA) was introduced in the 1980's as a model for Blind Source Separation (BSS), which refers to the process of recovering the sources underlying a mixture of signals, with little knowledge about the source signals or the mixing process. While there are many sophisticated algorithms for estimation, different methods have different shortcomings. In this paper, we develop a nonparametric score to adaptively pick the right algorithm for ICA with arbitrary Gaussian noise. The novelty of this score stems from the fact that it just assumes a finite second moment of the data and uses the characteristic function to evaluate the quality of the estimated mixing matrix without any knowledge of the parameters of the noise distribution. In addition, we propose some new contrast functions and algorithms that enjoy the same fast computability as existing algorithms like FASTICA and JADE but work in domains where the former may fail. While these also may have weaknesses, our proposed diagnostic, as shown by our simulations, can remedy them. Finally, we propose a theoretical framework to analyze the local and global convergence properties of our algorithms.

1 Introduction

Independent Component Analysis (ICA) was introduced in the 1980's as a model for Blind Source Separation (BSS) [13, 12], which refers to the process of recovering the sources underlying a mixture of signals, with little knowledge about the source signals or the mixing process. It has become a powerful alternative to the Gaussian model, leading to PCA in factor analysis. A good account of the history, many results, and the role of dimensionality in noiseless ICA can be found in [28, 5]. In one of its first forms, ICA is based on observing n independent samples from a k-dimensional population:

$$\boldsymbol{x} = B\boldsymbol{z} + \boldsymbol{g} \tag{1}$$

where $B \in \mathbb{R}^{d\times k}$ $(d \geq k)$ is an unknown mixing matrix, $\boldsymbol{z} \in \mathbb{R}^k$ is a vector of statistically independent mean zero "sources" or "factors" and $\boldsymbol{g} \sim N(\mathbf{0}, \Sigma)$ is a d dimensional mean zero Gaussian noise vector with covariance matrix Σ, independent of $\boldsymbol{z}$. We denote $\operatorname{diag}(\boldsymbol{a}) := \operatorname{cov}(\boldsymbol{z})$ and $S := \operatorname{cov}(\boldsymbol{x}) = B \operatorname{diag}(\boldsymbol{a}) B^T + \Sigma$. The goal of the problem is to estimate B.

ICA was initially developed in the context of $\boldsymbol{g} = 0$, now called noiseless ICA. In that form, it spawned an enormous literature [17, 39, 40, 1, 8, 27, 35, 37, 23, 11, 6, 42] and at least two well-established algorithms FASTICA [29] and JADE [9] which are widely used. The general model with nonzero $\boldsymbol{g}$ and unknown noise covariance matrix Σ, known as *noisy* ICA, has developed more slowly, with its own literature [4, 50, 49, 30, 31, 41].

In the following sections, we will show that various ICA and noisy ICA methods have distinct shortcomings. Some struggle with heavy-tailed distributions and outliers, while others require approximations of entropy-based objectives, which have their own challenges (see eg. [32]).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-4217

Although methods for noiseless cases may sometimes work in noisy settings [49], they are not always reliable (e.g., see Figure 1 in [49] and Theorem 2 in [11]). Despite the plethora of algorithms for noisy and noiseless ICA, the literature has largely been missing a diagnostic to decide which algorithm to pick for a given dataset. This is our primary goal.

The paper is organized as follows. Section 2 contains background and notation. Section 3.1 contains the independence score, its properties, and a Meta algorithm 1 that utilizes this score. Section 3.3 introduces new contrast functions based on the natural logarithm of the characteristic function and moment generating function of $\boldsymbol{x}^T\boldsymbol{u}$. Section 4 presents the global and local convergence results for noisy ICA. Section 5 demonstrates the empirical performance of our new contrast functions as well as the Meta algorithm applied to a variety of methods for inferring B.

2 Background and overview

This section contains a general overview of the main concepts of ICA.

Notation: We denote vectors using bold font as $\boldsymbol{v} \in \mathbb{R}^d$. The dataset is represented as $\left\{\boldsymbol{x}^{(i)}\right\}_{i\in[n]}$ for $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$. Scalar random variables and matrices are represented using upper case letters, respectively. $\boldsymbol{I}_k$ denotes the k- dimensional identity matrix. Entries of vector $\boldsymbol{v}$ (matrix M) are represented as v_i (M_{ij}). $M(:,i)$ ($M(i,:)$) represents the i^{th} column (row) of matrix M. $M^\dagger$ represents the Moore–Penrose inverse of M. $\|.\|$ represents the Euclidean l_2 norm for vectors and operator norm for matrices. $\|.\|_F$ similarly represents the Frobenius norm for matrices. $\mathbb{E}\left[.\right]$ ($\widehat{\mathbb{E}}\left[.\right]$) denotes the expectation (empirical average); $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$ denotes statistical independence. $\|.\|_{\psi_2}$ represents the subgaussian Orlicz norm (see Def. 2.5.6 in [46]). $X_n = O_P\left(a_n\right)$ implies that $\frac{X_n}{a_n}$ is stochastically bounded i.e, $\forall \epsilon > 0$, there exists a finite M such that $\sup_n \mathbb{P}\left(\left|\frac{X_n}{a_n}\right| > M\right) \leq \epsilon$ (See [45]).

Identifiability: One key issue in Eq 1 is the identifiability of $A := B^{-1}$[1], which holds up to permutation and scaling of the coordinates of $\boldsymbol{z}$ if, at most, one component of $\boldsymbol{z}$ is Gaussian (see [14, 43]). If $d = k$ and Σ are unknown, it is possible, as we shall see later, to identify the canonical versions of B and Σ. For $d > k$ and unknown Σ, it is possible (see [15]) to reduce to the $d = k$ case. In this work, we assume $d = k$. For classical factor models, when $\boldsymbol{z}$ is Gaussian, identifiability can only be identified up to rotation and scale, leading to non-interpretability. This suggests fitting strategies focusing on recovering B^{-1} through nongaussianity as well as independence. In the noisy model (Eq 1), an additional difficulty is that recovery of the source signals becomes impossible even if B is known exactly. This is because, the ICA models $\mathbf{x} = B(\mathbf{z} + \mathbf{r}) + \mathbf{g}$ and $\mathbf{x} = B\mathbf{z} + (B\mathbf{r} + \mathbf{g})$ are indistinguishable (see [48] pages 2-3). This is resolved in [48] via maximizing the Signal-to-Interference-plus-Noise ratio (SINR). In classical work on ICA, a large class of algorithms (see [37]) choose to optimize a measure of non-gaussianity, referred to as contrast functions.

Contrast functions: Methods for estimating B typically optimize an empirical contrast function. Formally, we define a contrast function $f(\boldsymbol{u}|P)$ by $f : \mathcal{B}_k \times \mathcal{P} \to \mathbb{R}$ where $\mathcal{B}_k$ is the unit ball in $\mathbb{R}^k$ and $\mathcal{P}$ is essentially the class of mean zero distributions on $\mathbb{R}^d$. More concretely, for suitable choices of g, $f(\boldsymbol{u}|P) \equiv f(\boldsymbol{u}|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}\sim P}[g(\boldsymbol{u}^T\boldsymbol{x})]$ for any random variable $\boldsymbol{x} \sim P$. The most popular example of a contrast function is the kurtosis, which is defined as $f(\boldsymbol{u}|P) = \mathbb{E}_{\boldsymbol{x}\sim P}\left(\boldsymbol{u}^T\boldsymbol{x}\right)^4 / \mathbb{E}\left[\left(\boldsymbol{u}^T\boldsymbol{x}\right)^2\right]^2 - 3$. The aim is to maximize an empirical estimate, $\hat{f}(\boldsymbol{u}|P)$. Notable contrast functions are the negentropy, mutual information (used by INFOMAX [7]), the $\tanh$ function, etc. (See [37] for further details).

Fitting strategies: There are two broad categories of the existing fitting strategies. (I) Find A such that the components of $A\boldsymbol{x}$ are both as nongaussian and as independent as possible. See, for example, the multivariate cumulant-based method JADE [9] and characteristic function-based methods [21, 11]. The latter we shall call the PFICA method. (II) Find successively the j^{th} row of A, denoted by $A(j,:) \in \mathbb{R}^d$, $j = 1, \ldots, k$ such that $A(j,:)^T\boldsymbol{x}$ are independent and each as nongaussian as possible, that is, estimate $A(1,:)$, project, and then apply the method again on the residuals successively. The chief representative of these methods is FastICA [27] based on univariate kurtosis.

From noiseless to noisy ICA: In the noiseless case one can first prewhiten the dataset using the empirical variance-covariance matrix, $\hat{\Sigma}$, of $\boldsymbol{x}^{(i)}$, using $\hat{\Sigma}^{-1/2}$. Therefore, WLOG, B can be assumed to be orthogonal. Searching over orthogonal matrices not only simplifies algorithm design for fitting

[1] If B is not invertible, this can be replaced by the Moore-Penrose inverse.

strategy (I) but also makes strategy (II) meaningful. It also greatly simplifies the analysis of kurtosis-based contrast functions (see [18]). For noisy ICA, prewhitening requires knowledge of the noise covariance, and therefore many elegant methods [4, 48, 49] avoid this step.

Individual shortcomings of different contrast functions and fitting strategies: To our knowledge, no single ICA algorithm or contrast function works universally across all source distributions. Adaptively determining which algorithm is useful in a given setting would be an important tool in a practitioner's toolbox. Key challenges include:

Cumulants: Even though contrast functions based on *even cumulants* have no spurious local optima (Theorem 3, [49]), they can vanish for some non-Gaussian distributions. Our experiments (Section 5) show that kurtosis-based contrast functions can perform poorly even when there are a few independent components with zero kurtosis. They also suffer under heavy-tailed source distributions [3, 11, 26].

PFICA: PFICA can outperform cumulant-based methods in some situations (Section 5). However, it is computationally much slower, and poorly performs for Bernoulli(p) sources with small p.

FastICA: [32] show that FastICA's use of Negentropy approximations for computational efficiency may result in a contrast function where optimal directions do not correspond to directions of low entropy. $\tanh$ is another popular smooth contrast function used by FastICA. However, in [52], the authors show that this tends to return spurious solutions for some distributions [52].

Contrast functions are usually nonconvex and may have spurious local optima. However, they may still, under suitable conditions, converge to maximal directions. We introduce such a contrast function which vanishes iff $\boldsymbol{x}$ is Gaussian in Section 3.3 and does not require higher moments.

Our contributions:

1) Using Theorem 1 we obtain a computationally efficient scoring method for evaluating the B estimated by different algorithms. Our score can also be extended to evaluate each demixing direction in sequential settings, which we do not pursue here.

2) We propose new contrast functions that work under scenarios where cumulant-based methods fail. We propose a fast sequential algorithm as in [49] to optimize these and further provide theoretical guarantees for convergence.

3 Main contributions

In this section, we present the "Independence score" and state the key result that justifies it. We conclude with two new contrast functions.

3.1 Independence score

While there are many tests for independence [22, 34, 42, 6], we choose the Characteristic function-based one because it is easy to compute and has been widely used in normality testing [19, 20, 25] and ICA [11, 21, 44]. Let us start with noiseless ICA to build intuition. Say we have estimated the inverse of the mixing matrix, i.e. B^{-1} using a matrix F. Ideally, if $F = B^{-1}$, then $F\boldsymbol{x} = \boldsymbol{z}$, where $\boldsymbol{z}$ is a vector of independent random variables. Kac's theorem [33] tells us that $\mathbb{E}\left[\exp(it^T\boldsymbol{z})\right] = \prod_{j=1}^{k}\mathbb{E}\left[\exp(it_jz_j)\right]$ if and only if z_i are independent. In [21], a novel characteristic function-based objective (CHFICA), is further analyzed and studied by [11] (PFICA). Here one minimizes $|\mathbb{E}\left[\exp(it^TF\boldsymbol{x})\right] - \prod_{j=1}^{k}\mathbb{E}\left[\exp(it_j(F\boldsymbol{x})_j)\right]|$ over F. We refer to this objective as the *uncorrected* independence score. We propose to adapt the CHFICA objective using *estimable parameters*, S, to the noisy ICA setting. We will minimize:

$$\Delta(\boldsymbol{t}, F|P) := \Bigg|\underbrace{\mathbb{E}\left[\exp(i\boldsymbol{t}^TF\boldsymbol{x})\right]\exp\left(\frac{-\boldsymbol{t}^T\operatorname{diag}(FSF^T)\boldsymbol{t}}{2}\right)}_{\text{JOINT}} - \underbrace{\prod_{j=1}^{k}\mathbb{E}\left[\exp(it_j(F\boldsymbol{x})_j)\right]\exp\left(\frac{-\boldsymbol{t}^TFSF^T\boldsymbol{t}}{2}\right)}_{\text{PRODUCT}}\Bigg| \tag{2}$$

where $S = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$ and can be estimated using the sample covariance matrix. Hence, we do not require knowledge of any model parameters. We refer to this score as the (*corrected*) independence score. The second terms in JOINT and PRODUCT (Eq 2) *corrects* the original score by canceling out the additional terms resulting from the Gaussian noise using the covariance matrix S of the

 https://doi.org/10.52202/079017-4217

data (estimated via the sample covariance). See [21] for a related score requiring knowledge of the unknown Gaussian covariance matrix Σ. We are now ready to present our first theoretical result for consistency of $\Delta(\boldsymbol{t}, F|P)$.

Theorem 1. *If $F \in \mathbb{R}^{k\times k}$ is invertible and the joint and marginal characteristic functions of all independent components, $\{z_i\}_{i\in[k]}$, are twice-differentiable, then $\forall \boldsymbol{t} \in \mathbb{R}^k$, $\Delta(\boldsymbol{t}, F|P) = 0$ iff $F = DB^{-1}$ where D is a permutation of an arbitrary diagonal matrix.*

Unfortunately, F is not uniquely defined if Σ is unknown as we have noted in Section 1. The proof of Theorem 1 is provided in the Appendix Section A.1. When using this score in practice, S is replaced with the sample covariance matrix of the data, and the expectations are replaced by the empirical estimates of the characteristic function. The convergence of this empirical score, $\Delta(\mathbf{t}, F|\hat{P})$, to the population score, $\Delta(\mathbf{t}, F|P)$, is given by the following theorem.

Theorem 2. *Let $\mathcal{F} := \{F \in \mathbb{R}^{k\times k} : \|F\| \leq 1\}$, $\boldsymbol{x} \sim$ subgaussian(σ) and $C_k := \max(1, k\log(n)\operatorname{Tr}(S))$. Then, we have*

$$\sup_{F\in\mathcal{F}} |\mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\Delta(\boldsymbol{t}, F|P) - \mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\Delta(\boldsymbol{t}, F|\hat{P})| = O_P\left(\sqrt{\frac{k^2\|S\|\max(k, \sigma^4\|S\|)\log^2(nC_k)}{n}}\right)$$

This bound shows that uniformly over F, the empirical average $\mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\Delta(\boldsymbol{t}, F|\hat{P})$, is close to the population score. This guarantees that as long as the difference between the population scores of the two candidate algorithms is not too small, the meta-algorithm can pick up the better score.

Remark 1 (Subgaussianity assumption)**.** *The subgaussianity assumption in Theorem 2 simply ensures the concentration of the sample covariance matrix since it is used in the score (see Theorem A.3, line 825). It can be relaxed if the concentration in the operator norm is ensured. Appendix Section A.2.1 contains experiments with more super-Gaussian source signals.*

The proof of this result is deferred to the Appendix section A.1.1. It is important to note that $\Delta(\boldsymbol{t}, F|\hat{P})$ is not easy to optimize. As we show in Section 3.2, objective functions that are computationally more amenable to optimize for ICA, e.g., cumulants, satisfy some properties (see Assumption 1). The independence score does not satisfy the first property (Assumption 1 (a)). In the *noiseless* case, whitening reduces the problem to B being orthogonal. This facilitates efficient gradient descent in the Stiefel manifold of orthogonal matrices [11]. However, in the noisy setting, the prewhitening matrix contains the unknown noise covariance Σ, making optimization hard. Furthermore, as noted in Remark 2, in practice, it is better to use $\mathbb{E}_{\boldsymbol{t}\sim N(0,I_k)}\Delta(\boldsymbol{t}, F|\hat{P})$ [11, 21, 25] instead of using a fixed vector, $\boldsymbol{t}$, to evaluate $\Delta(\boldsymbol{t}, F|\hat{P})$. Any iterative optimization method requires repeatedly estimating this expectation at each gradient step. Therefore, instead of using this score directly as the optimization objective, we use it to choose between different contrast functions after extracting the demixing matrix with each. This process is referred to as the Meta algorithm (Algorithm 1).

Algorithm 1 Meta-algorithm for choosing best candidate algorithm.

Input: Algorithm list $\mathcal{L}$, Data $X \in \mathbb{R}^{n\times k}$
for j in range[1, size $(\mathcal{L})$] **do**
  $B_j \leftarrow \mathcal{L}_j(X)$ {Extract mixing matrix B_j using j^{th} candidate algorithm}
  $\delta_j \leftarrow \widehat{\mathbb{E}}_{\boldsymbol{t}\sim\mathcal{N}(0,\boldsymbol{I}_k)}\left[\Delta\left(\mathbf{t}, B_j^{-1}|\hat{P}\right)\right]$
end for
$i_* \leftarrow \arg\min_{j\in[\text{size}(\mathcal{L})]}[\delta_j]$
return B_{i_*}

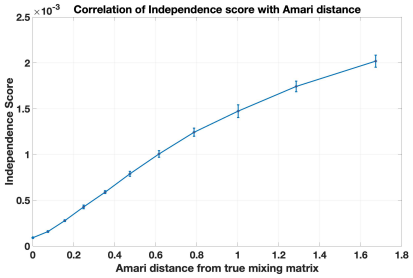


Figure 1: Correlation of independence score (with std. dev.) with Amari error between $B' = \epsilon B + (1-\epsilon) I$ and B, averaged over 10 random runs.

Remark 2 (Averaging over $\boldsymbol{t}$)**.** *Algorithm 1 averages the independence score over $\boldsymbol{t} \sim \mathcal{N}(0, \boldsymbol{I}_k)$. While Eq 2 defines the independence score for one direction $\boldsymbol{t}$, in practice, however, there may be some directions such that a non-gaussian signal has a small score in that direction. Hence, it is desirable to average over $\boldsymbol{t}$, following the convention in [11, 21, 25].*

To gain an intuition of the independence score, we conduct an experiment where we use the dataset mentioned in Section 5 (Figure 2b) and compute the independence score for $B' = \epsilon B + (1-\epsilon)\, I, \epsilon \in (0.5, 1)$. As we increase ϵ, B' approaches B, and hence the Amari error $d_{B',B}$ (see Eq 8) decreases. Figure 1 shows the independence score versus Amari error, indicating that the independence score accurately predicts solution quality even without knowing the true B.

3.2 Desired properties of contrast functions

The following properties are often useful for designing provable optimization algorithms for ICA.

Assumption 1 (Properties of contrast functions)**.** *Let f be a contrast function defined for a mean zero random variable X. Then,*

(a) $f(u|X+Y) = f(u|X) + f(u|Y)$ *for* $X \perp\!\!\!\perp Y$

(b) $f(0|X) = 0, f'(0|X) = 0,\ f''(0|X) = 0$

(c) $f(u|G) = 0, \forall u$ *for* $G \sim \mathcal{N}(0,1)$

(d) WLOG, $f(u|X) = f(-u|X)$ *(symmetry)*

Properties (a) and (c) ensure that $f(u|G+X) = f(u|X)$ for non-gaussian X and independent non-gaussian G, which means that the additive independent Gaussian noise does not change f. Property (c) ensures that $|f(u|G)|$ is minimal for a Gaussian. Property (d) holds without loss of generality because one can always symmetrize the distribution. [2]

3.3 New contrast functions

In Section 2, we provided a discussion of the individual shortcomings of different contrast functions for existing contrast functions. Before we introduce new contrast functions in this section, we revisit the algorithmic issues posed by the added Gaussian noise with unknown Σ in Eq 1.

Prewhitening the data is challenging for noisy ICA because $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$ includes the unknown Gaussian covariance matrix Σ. [4] show that it is enough to *quasi* orthogonalize the data, i.e., multiply it by a matrix, which makes the data have a diagonal covariance matrix (not necessarily the identity). Subsequent work [50, 49] uses cumulants and the Hessian of $f(\boldsymbol{u})$ to construct a matrix $C := BDB^T$ where D is a diagonal matrix, and then use this matrix to achieve quasi-orthogonalization. In general, D may not be positive semidefinite (e.g. when components of $\boldsymbol{z}$ have kurtosis of different signs). To remedy this, in [49], the authors propose computing the C matrix using the Hessian of $f(\boldsymbol{u})$ at some fixed unit vector and then perform an elegant power method in the pseudo-Euclidean space:

$$\boldsymbol{u}_t \leftarrow \nabla f(C^\dagger \boldsymbol{u}_{t-1}|P) \big/ \|\nabla f(C^\dagger \boldsymbol{u}_{t-1}|P)\| \tag{3}$$

A pseudo-Euclidean space is a generalization of the Euclidean space used by [48]. Here the produce between vectors $\boldsymbol{u}, \boldsymbol{v}$ is given as $u^\top A v$,

Algorithm 2 Pseudo-Euclidean Power Iteration for ICA (Algorithm 2 in [48]) $\tilde{B}$ is the recovered matrix for the ICA model in Eq 1. $\tilde{A}$ is a running estimate of $\tilde{B}^\dagger$.

Input: $C \in \mathbb{R}^{k \times k}$, ∇f
$\tilde{A} \leftarrow 0$, $\tilde{B} \leftarrow 0$
for j in range[1, k] **do**
  Draw $\mathbf{u}$ uniformly at random from $\mathcal{S}^{k-1}$
  while Convergence (up to sign) **do**
    $\mathbf{u} \leftarrow \tilde{B}\tilde{A}\mathbf{u}$
    $\mathbf{u} \leftarrow \nabla f\left(C^\dagger \mathbf{u}\right) / \left\|\nabla f\left(C^\dagger \mathbf{u}\right)\right\|_2$
  end while
  $\tilde{B}_j \leftarrow \mathbf{u}$, $\tilde{A}_j \leftarrow \left[C^\dagger \tilde{B}_j \Big/ \left(\left(C^\dagger \tilde{B}_j\right)^\top \tilde{B}_j\right)\right]^\top$
end for
return $\tilde{B}, \tilde{A}$

[2] Note that $y_i := x_i - x_{\lfloor n/2 \rfloor + i},\ i = 1 \cdots \lfloor n/2 \rfloor$ has a symmetric distribution, and also remains a noisy version of a linear combination of source signals with the same mixing matrix.

https://doi.org/10.52202/079017-4217

In this section, we present two new contrast functions based on the logarithm of the characteristic function (CHF) and the cumulant generating function (CGF). These do not depend on a particular cumulant, and the first only requires a finite second moment, which makes it suitable for heavy-tailed source signals. We will use $f(\boldsymbol{u}|P)$ or $f(\boldsymbol{u}|\boldsymbol{x})$ where $\boldsymbol{x} \sim P$ interchangeably to represent the population contrast function. In constructing both of the following contrast functions, we use the fact that, like the cumulants, the CGF and CHF-based contrast functions satisfy Assumption 1 (a). To satisfy Assumption 1 (c), we subtract out the part resulting from the Gaussian noise, which leads to the additional terms involving $\boldsymbol{u}^\top S\boldsymbol{u}$.

CHF-based contrast function: Recall that S denotes the covariance matrix of $\boldsymbol{x}$. We maximize the *absolute value* of following:

$$\begin{aligned} f(\boldsymbol{u}|P) &= \log \mathbb{E} \exp(i\boldsymbol{u}^T\boldsymbol{x}) + \log \mathbb{E} \exp(-i\boldsymbol{u}^T\boldsymbol{x}) + \boldsymbol{u}^T S\boldsymbol{u} \\ &= \log(\mathbb{E} \cos(\boldsymbol{u}^T\boldsymbol{x}))^2 + \log(\mathbb{E} \sin(\boldsymbol{u}^T\boldsymbol{x}))^2 + \boldsymbol{u}^T S\boldsymbol{u} \end{aligned} \tag{4}$$

The intuition is that this is exactly zero for zero mean Gaussian data and maximizing this leads to extracting non-gaussian signal from the data. It also satisfies Assumption 1 a), b) and c).

CGF-based contrast function: The cumulant generating function also has similar properties (see [53] for the noiseless case). In this case, we maximize the *absolute value* of following:

$$f(\boldsymbol{u}|P) = \log \mathbb{E} \exp(\boldsymbol{u}^T\boldsymbol{x}) - \frac{1}{2}\boldsymbol{u}^T S\boldsymbol{u} \tag{5}$$

Like CHF, this vanishes iff $\boldsymbol{x}$ is mean zero Gaussian, and satisfies Assumption 1 a), b) and c). However, it cannot be expected to behave well in the heavy-tailed case. In the Appendix (Section A.1, Theorem A.1), we show that the Hessian of both functions obtained from Eq 4 and Eq 5 evaluated at some $\boldsymbol{u}$ is, in fact, of the form BDB^T where D is some diagonal matrix.

4 Global and local Convergence: loss landscape of noisy ICA

Here we present sufficient conditions for global and local convergence of a broad class of contrast functions. This covers the cumulant-based contrast functions and our CHF and CGF-based proposals.

4.1 Global convergence

The classical argument for global optima of kurtosis for noiseless ICA in [18] assumes WLOG that B is orthogonal. This reduces the optimization over $\boldsymbol{u}$ into one for $\boldsymbol{v} = B\boldsymbol{u}$. Since B is orthogonal, $\|\boldsymbol{u}\| = 1$ translates into $\|\boldsymbol{v}\| = 1$. It may seem that this idea should extend to the noisy case due to property 1 a) and c) by optimizing over $\boldsymbol{v} = B\boldsymbol{u}$ over potentially non-orthogonal mixing matrices B.

However, for non-orthogonal B, the norm constraint on $\boldsymbol{v}$ is no longer $\boldsymbol{v}^T\boldsymbol{v} = 1$, *rendering the original argument invalid.* In what follows, we extend the original argument to the noisy ICA setting by introducing pseudo-euclidean constraints. Our framework includes a large family of contrast functions, including cumulants.

To our knowledge, this is the first work to characterize the loss functions in the pseudo-euclidean space. In contrast, [50] provides global convergence of the cumulant-based methods by a convergence argument of the power method itself.

Consider the contrast function $f(\boldsymbol{u}|P) := \mathbb{E}_{\boldsymbol{x}\sim P}\left[g\left(\boldsymbol{x}^T\boldsymbol{u}\right)\right]$ and recall the definition of the quasi orthogonalization matrix $C = BDB^T$, where D is a diagonal matrix. For simplicity, WLOG let us assume that D is invertible. We now aim to find the optimization objective that leads to the pseudo-Euclidean update in Eq 3. For $f\left(C^{-1}\boldsymbol{u}|P\right) = \mathbb{E}_{\boldsymbol{x}\sim P}\left[g\left(\boldsymbol{u}^T C^{-1}\boldsymbol{x}\right)\right]$, consider the following:

$$f\left(C^{-1}\boldsymbol{u}|P\right) = \sum_{i\in[k]} \mathbb{E}\left[g\left(\left(B^{-1}\boldsymbol{u}\right)_i z_i/D_{ii}\right)\right] = \sum_{i\in[k]} \mathbb{E}\left[g\left(\alpha_i \tilde{z}_i\right)\right] = \sum_{i\in[k]} f\left(\alpha_i/D_{ii}|z_i\right) \tag{6}$$

where we define $\boldsymbol{\alpha} := B^{-1}\boldsymbol{u}$ and $\tilde{z}_i = z_i/D_{ii}$. We now examine $f(C^{-1}\boldsymbol{u})$ subject to the "pseudo" norm constraint $\boldsymbol{u}^T C^{-1}\boldsymbol{u} = 1$. The key point is that for suitably defined f, one can construct a matrix $C = BDB^T$ from the data even when B is unknown. So our new objective is to optimize

$$f\left(C^{-1}\boldsymbol{u}|P\right) \text{ s.t. } \boldsymbol{u}^T C^{-1}\boldsymbol{u} = 1$$

Using the Lagrange multiplier method, optimizing the above simply gives the power method update (Eq 3) $\boldsymbol{u} \propto \nabla f(C^{-1}\boldsymbol{u})$. Furthermore, optimizing Eq 6 leads to the following transformed objective:

$$\max_{\boldsymbol{\alpha}} \left| \sum_i f\left(\alpha_i / D_{ii} | z_i\right) \right| \quad \text{s.t.} \quad \sum_i \alpha_i^2 / D_{ii} = 1 \tag{7}$$

Now we provide our theorem about maximizing $f(C^{-1}\boldsymbol{u}|P)$. Analogous results hold for minimization. We will denote the corresponding derivatives evaluated at t_0 as $f'(t_0|z_i)$ and $f''(t_0|z_i)$.

Theorem 3. *Let C be a matrix of the form BDB^T, where $d_i := D_{ii} = f''(u_i|z_i)$ for some random u_i. Let $S_+ = \{i : d_i > 0\}$. Consider a contrast function $f : \mathbb{R} \to \mathbb{R}$. Assume that for every non-Gaussian independent component, Assumption 1 holds and the third derivative, $f'''(u|X)$, does not change the sign in the half-lines $[0, \infty)$, $(\infty, 0]$. Then $f\left(C^{-1}\boldsymbol{u} \middle| \boldsymbol{x}\right)$ with the constraint of $\langle \boldsymbol{u}, \boldsymbol{u} \rangle_{C^{-1}} = 1$ has local maxima at $B^{-1}\boldsymbol{u} = \boldsymbol{e}_i, i \in S_+$. All solutions satisfying $\boldsymbol{e}_i^T B^{-1}\boldsymbol{u} \neq 0, \ \ \forall i \in [1, k]$ are minima, and all solutions with $|\left\{i : \boldsymbol{e}_i^T B^{-1}\boldsymbol{u} \neq 0\right\}| < k$ are saddle points. These are the only fixed points.*

Note that the above result immediately implies that for $\tilde{C} = -C$, the same optimization in Theorem 3 will have maxima at all $\boldsymbol{u}$ such that $B^{-1}\boldsymbol{u} = \boldsymbol{e}_i, i \in S_-$.

Corollary 3.1. *$2k^{th}$ cumulant-based contrast functions for $k > 1$, have maxima and minima at the directions of the columns of B, provided $\{\boldsymbol{z}_i\}_{i \in [k]}$ have a corresponding non-zero cumulant.*

Proof. This proof (see Appendix Section A.1) is immediate since cumulants satisfy (a), (c). For (c) and (d), note that $f(u|Z)$ is simply $u^{2j}\kappa_{2j}(Z)$, where $\kappa_{2j}(Z)$ is the $2j^{th}$ cumulant of Z. □

Theorem 3 can be easily extended to the case where C is not invertible. The condition on the third derivative, $f'''(u|X)$, may seem strong, and it is not hard to construct examples of random variables Z where this fails for suitable contrast functions (see [36]). However, for such settings, it is hard to separate Z from a Gaussian. See the Appendix A.4.1 for more details.

4.2 Local convergence

In this section, we establish a local convergence result for the power iteration-based method described in Eq 3. Define $\forall t \in \mathbb{R}, \forall i \in [d]$, the functions $q_i(t) := f'\left(\frac{\alpha_i}{D_{ii}} \middle| z_i\right)$. Then, without loss of generality, we have the following result for the first corner:

Theorem 4. *Let $\alpha_i = B^{-1}\boldsymbol{u}_i$ and $\boldsymbol{\alpha}^* := \boldsymbol{e}_1$. Let the contrast function $f(.|X)$, satisfy assumptions 1 (a), (b), (c), and (d). Let $\mathcal{B} := [-\|B^{-1}\|_2, \|B^{-1}\|_2]$ and define $c_1, c_2, c_3 > 0$ such that $\forall i \in [d]$, $\sup_{x \in \mathcal{B}} \left\{ \frac{|q_i(x)|}{c_1}, \frac{|q_i'(x)|}{c_2}, \frac{|q_i''(x)|}{c_3} \right\} \leq 1$. Define $\epsilon := \frac{\|B\|_F}{\|B\mathbf{e}_1\|^2} \max\left\{ \frac{c_3 + c_2\|B\mathbf{e}_1\|}{|q_1(\alpha_1^*)|}, \frac{c_2^2}{|q_1(\alpha_1^*)|^2}, \frac{c_3\|B\mathbf{e}_1\|}{|q_1'(\alpha_1^*)|} \right\}$. Let $\|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*\|_2 \leq R$, $R \leq \max\left\{\frac{c_2}{c_3}, \frac{1}{5\epsilon}\right\}$. Then, we have $\forall t \geq 1$, $\frac{\|\boldsymbol{\alpha}_{t+1} - \boldsymbol{\alpha}^*\|}{\|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*\|} \leq \frac{1}{2}$.*

Therefore, we can establish linear convergence without the third derivative constraint required for global convergence (see Theorem 3), by assuming some mild regularity conditions on the contrast functional and a sufficiently close initialization. The proof is included in Appendix Section A.1.2.

Remark 3. *Let $\boldsymbol{\alpha} = B^{-1}\boldsymbol{u}$ denote the demixed direction, $\boldsymbol{u}$. Theorem 4 shows that if the initial $\boldsymbol{u}_0$ is such that $\boldsymbol{\alpha}_0$ is close to the ground truth $\boldsymbol{\alpha}^*$ (which is WLOG taken to be $\boldsymbol{e}_1$), then there is geometric convergence to $\boldsymbol{\alpha}^*$. $|q_1(\alpha_1^*)|$ and $|q_1'(\alpha_1^*)|$ essentially quantify how non-gaussian the corresponding independent component is since they are zero for a mean zero Gaussian. When these are large quantities, ϵ is small, and the initialization radius $\|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*\|_2$ can be relatively larger.*

5 Experiments

In this section, we provide experiments to compare the fixed-point algorithms based on the characteristic function (CHF), the cumulant generating function (CGF) (Eqs 4 and 5) with the kurtosis-based algorithm (PEGI-κ_4 [49]). We also compare against noiseless ICA[3] algorithms - FastICA, JADE, and PFICA. These six algorithms are included as candidates for the Meta algorithm (see Algorithm

[3]MATLAB implementations (under the GNU General Public License) can be found at - FastICA and JADE. The code for PFICA was provided on request by the authors.

 https://doi.org/10.52202/079017-4217

1). We also present the *Uncorrected Meta* algorithm (*Unc. Meta*) to denote the Meta algorithm with the uncorrected independence score proposed in CHFICA [21]. Table 1 and Figure 2 provide experiments on simulated data, and Figure 3 provides an application for image demixing [16]. We provide additional experiments on denoising MNIST images in the Appendix Section A.2.

Experimental setup: Similar to [49], the mixing matrix B is constructed as $B = U\Lambda V^T$, where U, V are k dimensional random orthonormal matrices, and Λ is a diagonal matrix with $\Lambda_{ii} \in [1, 3]$. The covariance matrix Σ of the noise $\boldsymbol{g}$ follows the Wishart distribution and is chosen to be $\frac{\rho}{k} RR^T$, where k is the number of sources, and R is a random Gaussian matrix. Higher values of the *noise power* ρ can make the noisy ICA problem harder. Keeping B fixed, we report the median of 100 random runs of data generated from a given distribution (different for different experiments). The quasi-orthogonalization matrix for CHF and CGF is initialized as $\hat{B}\hat{B}^T$ using the mixing matrix, $\hat{B}$, estimated via PFICA. The performance of CHF and CGF is based on a single random initialization of the vector in the power method (see Algorithm 2). Our experiments were performed on a Macbook Pro M2 2022 CPU with 8 GB RAM.

Error Metrics: Due to the inherent ambiguity in signal recovery discussed in Section 2, we measure error using the accuracy of estimating B. We report the widely used Amari error [2, 24, 11, 10] for our results. For estimated demixing matrix $\widehat{B}$, define $W = \widehat{B}^{-1}B$, after normalizing the rows of $\widehat{B}^{-1}$ and B^{-1}. Then we measure $d_{\widehat{B},B}$ as -

$$d_{\widehat{B},B} := \frac{1}{k}\left(\sum_{i=1}^{k}\frac{\sum_{j=1}^{k}|W_{ij}|}{\max_j|W_{ij}|} + \sum_{j=1}^{k}\frac{\sum_{i=1}^{k}|W_{ij}|}{\max_i|W_{ij}|}\right) - 2 \tag{8}$$

Variance reduction using Meta: We first show that the independence score can also be used to pick the best solution from many random initializations. In Figure 2 (a), the top panel has a histogram of 40 runs of CHF, each with one random initialization. The bottom panel shows the histogram of 40 experiments, where, for each experiment, the *best out of 30 random initializations* are picked using the independence score. The top and bottom panels have (mean, standard deviation) (0.51, 0.51) and (0.39,0.34) respectively. This shows a reduction in variance and overall better performance.

Effect of varying kurtosis: For our second experiment (see Table 1), we use $k = 5$ independent components, sample size $n = 10^5$ and noise power $\rho = 0.2$ from a Bernoulli(p) distribution, where we vary p from 0.001 to $0.5 - 1/\sqrt{12}$. The last parameter makes kurtosis zero. Different algorithms

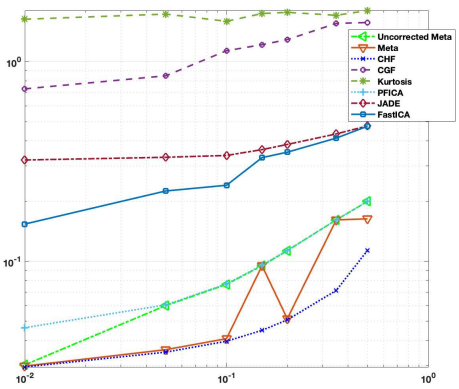
Scaled κ_4 / Algorithm	994	194	95	15	5	2	0.8	0.13	0
Meta	**0.007**	**0.010**	**0.011**	**0.010**	**0.011**	**0.011**	**0.0128**	**0.01981**	**0.023**
CHF	1.524	0.336	**0.011**	**0.010**	**0.011**	**0.011**	**0.0129**	**0.0213**	0.029
CGF	**0.007**	0.011	**0.011**	0.016	0.029	0.044	0.05779	0.06521	0.071
PEGI	**0.007**	**0.010**	**0.011**	**0.010**	0.012	0.017	0.02795	0.13097	1.802
PFICA	1.525	0.885	0.540	0.024	0.023	0.023	0.0212	0.0224	**0.024**
JADE	0.021	0.022	0.021	0.022	0.022	0.023	0.029	0.089	1.909
FastICA	0.024	0.027	0.026	0.026	0.026	0.027	0.02874	0.0703	-
Unc. Meta	1.52	0.049	0.041	0.0408	0.0413	0.0414	0.0413	0.0416	0.0419

Table 1: Median Amari error with varying p in Bernoulli(p) data. The scaled kurtosis is given as $\kappa_4 := (1 - 6p\,(1-p))/(p\,(1-p))$. Observe that the Meta algorithm (shaded in red) performs at par or better than the best candidate algorithms. FastICA did not converge for zero-kurtosis data.

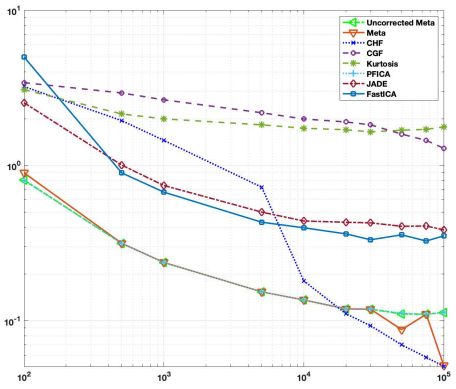
perform differently (best candidate algorithm is highlighted in bold font) for different p. In particular, characteristic function-based methods like PFICA and CHF perform poorly for small values of p. We attribute this to the fact that the characteristic function is close to one for small p. Kurtosis-based algorithms like PEGI, JADE, and FASTICA perform poorly for kurtosis close to zero. Furthermore, the Uncorrected Meta algorithm performs worse than the Meta algorithm since it shadows PFICA.

Effect of varying noise power: For our next two experiments, we use $k = 9$, out of which 3 are from Uniform $\left(-\sqrt{3}, \sqrt{3}\right)$, 3 are from Exponential(5) and 3 are from $\left(\text{Bernoulli}\left(\frac{1}{2} - \sqrt{\frac{1}{12}}\right)\right)$ and hence have zero kurtosis. In these experiments, we vary the sample size (Figure 2b), n fixing $\rho = 0.2$, and

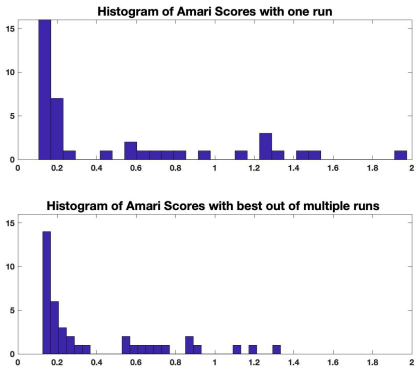
the noise power (Figure 2a) , ρ fixing $n = 10^5$, respectively. Note that with most mixture distributions, it is easily possible to have low or zero kurtosis. We include such signals in our data to highlight some limitations of PEGI-κ_4 and show that Algorithm 1 can choose adaptively to obtain better results. Figure 2a shows that large noise power leads to worse performance for all methods. PEGI, JADE,



(a) Variation of performance with noise power



(b) Variation of performance with sample size



(c) Histograms of Amari error with $n = 10^4$ and noise power $\rho = 0.2$

Figure 2: Amari error in the log-scale on y axis and varying noise powers (for $n = 10^5$) and varying sample sizes (for $\rho = 0.2$) on x axis for figures 2a and 2b respectively. For figure 2c, the top panel contains a histogram of 40 runs with one random initialization. The bottom panel contains a histogram of 40 runs, each of which is the best independence score out of 30 random initializations.

and FastICA perform poorly consistently, because of some zero kurtosis components. CGF suffers because of heavy-tailed components. However, CHF and PFICA perform well consistently. The Meta algorithm mostly picks the best algorithm except for a few points, where the difference between the two leading algorithms is small. The Uncorrected Meta algorithm (which uses the independence score without the additional correction term introduced for noise) performs identically to PFICA.

Effect of varying sample size: Figure 2b shows that the noiseless ICA methods have good performance at smaller sample sizes. However, they suffer a bias in performance compared to their noiseless counterparts as the sample size increases. The Meta algorithm can consistently pick the best option amongst the candidates, irrespective of the distribution, leading to significant improvements in performance. What is interesting is that up to a sample size of 10^4, PFICA dominates other algorithms and Meta performs like PFICA. However, after that CHF dominates, and one can see that Meta starts to have a similar trend as CHF. We also see that the Uncorrected Meta algorithm (which uses the independence score without the additional correction term introduced for noise) has a near identical error as PFICA and has a bias for large n.

6 Conclusion

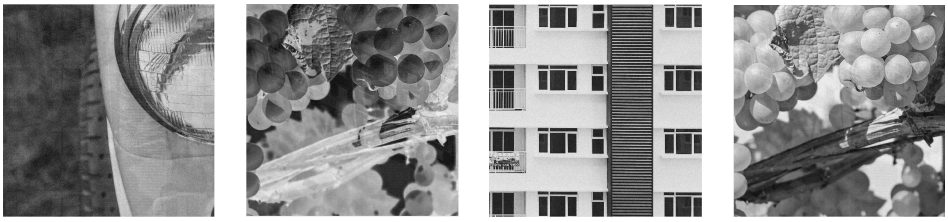
ICA is a classical problem that aims to extract independent non-Gaussian components. The vast literature on both noiseless and noisy ICA introduces many different inference methods based on a

https://doi.org/10.52202/079017-4217

(a) Source Images

(b) Mixed Images

(c) Demixed Images using CHF-based contrast function ($\Delta(\boldsymbol{t}, F|P) = 5.26 \times 10^{-3}$).



(d) Demixed Images using Kurtosis-based contrast function ($\Delta(\boldsymbol{t}, F|P) = 2.48 \times 10^{-2}$)

Figure 3: We demix images using ICA by flattening and linearly mixing them with a 4×4 matrix B (i.i.d entries $\sim \mathcal{N}(0, 1)$) and Wishart noise ($\rho = 0.001$). The CHF-based method (c) recovers the original sources well, upto sign. The Kurtosis-based method (d) fails to recover the second source. This is consistent with its higher independence score. The Meta algorithm selects CHF from candidates CHF, CGF, Kurtosis, FastICA, and JADE. Appendix Section A.2 provides results for other contrast functions and their independence scores.

variety of contrast functions for separating the non-Gaussian signal from the Gaussian noise. Each has its own set of shortcomings. We aim to identify the best method for a given dataset in a data-driven fashion. In this paper, we propose a nonparametric score, which is used to evaluate the quality of the solution of any inference method for the noisy ICA model. Using this score, we design a Meta algorithm, which chooses among a set of candidate solutions of the demixing matrix. We also provide new contrast functions and a computationally efficient optimization framework. While they also have shortcomings, we show that our diagnostic can remedy them. We provide uniform convergence properties of our score and theoretical results for local and global convergence of our methods. Simulated and real-world experiments show that our Meta-algorithm matches the accuracy of the best candidate across various distributional settings.

Acknowledgments and Disclosure of Funding

The authors thank Aiyou Chen for many important discussions that helped shape this paper and for sharing his code of PFICA. The authors also thank Joe Neeman for sharing his valuable insights and the anonymous reviewers for their helpful suggestions in improving the exposition of the paper. SK and PS were partially supported by NSF grants 2217069, 2019844, and DMS 2109155.

References

[1] Shun-Ichi Amari and J.-F. Cardoso. Blind source separation-semiparametric statistical approach. *IEEE Transactions on Signal Processing*, 45(11):2692–2700, 1997.

[2] Shun-ichi Amari, Andrzej Cichocki, and Howard Yang. A new learning algorithm for blind signal separation. *Advances in neural information processing systems*, 8, 1995.

[3] Joseph Anderson, Navin Goyal, Anupama Nandi, and Luis Rademacher. Heavy-tailed analogues of the covariance matrix for ica. In Satinder P. Singh and Shaul Markovitch, editors, *AAAI*, pages 1712–1718. AAAI Press, 2017.

[4] Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. "provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders". In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 2384–2392, 2012.

[5] Arnab Auddy and Ming Yuan. Large dimensional independent component analysis: Statistical optimality and computational tractability, 2023.

[6] F.R. Bach and M.I. Jordan. Kernel independent component analysis. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 4, pages IV–876, 2003.

[7] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[8] Jean-François Cardoso. High-order contrasts for independent component analysis. *Neural computation*, 11(1):157–192, 1999.

[9] J.F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings F (Radar and Signal Processing)*, 140:362–370(8), December 1993.

[10] Aiyou Chen and Peter J. Bickel. Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825 – 2855, 2006.

[11] Aiyou Chen and P.J. Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53(10):3625–3632, 2005.

[12] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[13] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[14] George Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pages 2–8, 1953.

[15] M. Davies. Identifiability issues in noisy ica. *IEEE Signal Processing Letters*, 11(5):470–473, 2004.

[16] Laurent de Vito. Ica for demixing images. `https://github.com/ldv1/ICA_for_demixing_images`, 2024. Accessed: 2024-05-20.

[17] N. Delfosse and P. Loubaton. Adaptive separation of independent sources: a deflation approach. In *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume iv, pages IV/41–IV/44 vol.4, 1994.

[18] Nathalie Delfosse and Philippe Loubaton. Adaptive blind separation of convolutive mixtures. In *ICASSP*, pages 2940–2943. IEEE Computer Society, 1996.

[19] Bruno Ebner and Norbert Henze. Bahadur efficiencies of the epps–pulley test for normality. *arXiv preprint arXiv:2106.13962*, 2021.

[20] Bruno Ebner and Norbert Henze. On the eigenvalues associated with the limit null distribution of the epps-pulley test of normality. *Statistical Papers*, 64(3):739–752, 2023.

 https://doi.org/10.52202/079017-4217

[21] Jan Eriksson and Visa Koivunen. Characteristic-function-based independent component analysis. *Signal Processing*, 83(10):2195–2208, 2003.

[22] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J. Smola. A kernel statistical test of independence. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 585–592, Red Hook, NY, USA, 2007. Curran Associates Inc.

[23] Trevor Hastie and Rob Tibshirani. Independent components analysis through product density estimation. *Advances in neural information processing systems*, 15, 2002.

[24] Trevor Hastie and Rob Tibshirani. Independent components analysis through product density estimation. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, page 665–672, Cambridge, MA, USA, 2002. MIT Press.

[25] Norbert Henze and Thorsten Wagner. A new approach to the bhep tests for multivariate normality. *Journal of Multivariate Analysis*, 62(1):1–23, 1997.

[26] A. Hyvarinen. One-unit contrast functions for independent component analysis: a statistical analysis. In *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*, pages 388–397, 1997.

[27] Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

[28] Aapo Hyvärinen, J. Karhunen, and Erkki Oja. *Independent component analysis*. John Wiley & Sons, 2001.

[29] Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[30] Pauliina Ilmonen and Davy Paindaveine. Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *Annals of Statistics*, 39:2448–2476, 2011.

[31] Pauliina Ilmonen and Davy Paindaveine. Semiparametrically efficient inference based on signed ranks in symmetric independent component models. 2011.

[32] Elena Issoglio, Paul Smith, and Jochen Voss. On the estimation of entropy in the fastica algorithm. *Journal of Multivariate Analysis*, 181:104689, 2021.

[33] M. Kac. Sur les fonctions indépendantes (i) (propriétés générales). *Studia Mathematica*, 6:46–58, 1936.

[34] Sergey Kirshner and Barnabás Póczos. Ica and isa using schweizer-wolff measure of dependence. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 464–471, New York, NY, USA, 2008. Association for Computing Machinery.

[35] Te-Won Lee, Mark Girolami, and Terrence J Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation*, 11(2):417–441, 1999.

[36] Peter McCullagh. Does the moment-generating function characterize a distribution? *The American Statistician*, 48(3):208–208, 1994.

[37] Erkki Oja and A Hyvarinen. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[38] Erkki Oja, Aapo Hyvärinen, and Patrik Hoyer. Image feature extraction and denoising by sparse coding. *Pattern Analysis & Applications*, 2:104–110, 1999.

[39] Dinh Tuan Pham. Blind separation of instantaneous mixture of sources via an independent component analysis. *IEEE Transactions on Signal Processing*, 44(11):2768–2779, 1996.

[40] Dinh Tuan Pham and P. Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Transactions on Signal Processing*, 45(7):1712–1725, 1997.

[41] Karl Rohe and Muzhe Zeng. Vintage factor analysis with varimax performs statistical inference. *arXiv preprint arXiv:2004.05387*, 2020.

[42] Hao Shen, Stefanie Jegelka, and Arthur Gretton. Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, 57(9):3498–3511, 2009.

[43] Viktor P Skitovitch. On a property of the normal distribution. *DAN SSSR*, 89:217–219, 1953.

[44] Fabian J. Theis. Multidimensional independent component analysis using characteristic functions. In *2005 13th European Signal Processing Conference*, pages 1–4, 2005.

[45] Aad W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[46] Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, Cambridge, UK, 2018.

[47] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[48] James R. Voss, Mikhail Belkin, and Luis Rademacher. Optimal recovery in noisy ICA. *CoRR*, abs/1502.04148, 2015.

[49] James R Voss, Mikhail Belkin, and Luis Rademacher. A pseudo-euclidean iteration for optimal recovery in noisy ica. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[50] James R Voss, Luis Rademacher, and Mikhail Belkin. Fast algorithms for gaussian noise invariant independent component analysis. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[51] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[52] Tianwen Wei. A study of the fixed points and spurious solutions of the deflation-based fastica algorithm. *Neural Comput. Appl.*, 28(1):13–24, jan 2017.

[53] Arie Yeredor. Blind source separation via the second characteristic function. *Signal Processing*, 80(5):897–902, 2000.

 https://doi.org/10.52202/079017-4217

A Appendix

The Appendix is organized as follows -

- Section A.1 proves Theorems 1, 2, A.1, 3 and 4
- Section A.2 provides additional experiments for noisy ICA
- Section A.3 provides the algorithm to compute independence scores via a sequential procedure
- Section A.4 explores the third derivative constraint in Theorem 3 and provides examples where it holds
- Section A.5 contains surface plots of the loss landscape for noisy ICA using the CHF-based contrast functions

A.1 Proofs

Proof of Theorem 1. We will give an intuitive argument for the easy direction for $k = 2$. We will show that if $F = DB^{-1}$ for some permutation of a diagonal matrix D, then after projecting using that matrix, the data is of the form $\boldsymbol{z} + \boldsymbol{g}'$ for some Gaussian vector $\boldsymbol{g}'$. Let $k = 2$. Let the entries of this vector be $z_1 + g_1'$ and $z_2 + g_2'$, where z_i, $i \in \{1, 2\}$ are mean zero *independent* non-Gaussian random variables and g_i', $i \in \{1, 2\}$ are mean zero *possibly dependent* Gaussian variables. The Gaussian variables are independent of the non-Gaussian random variables. Let t_i, $i \in \{1, 2\}$ be arbitrary but fixed real numbers. Let $\boldsymbol{t} = (t_1, t_2)$ and let $\Sigma_{g'}$ denote the covariance matrix of the Gaussian g'. Assume for simplicity $\text{var}(z_i) = 1$ for $i \in \{1, 2\}$. Denote by $\Lambda := \text{cov}(F\boldsymbol{x}) = FSF^T = I + \Sigma_{g'}$. The JOINT part of the score is given by:

$$\text{JOINT} = \mathbb{E}\left[it_1 z_1\right] \mathbb{E}\left[it_2 z_2\right] \exp\left(-\frac{\boldsymbol{t}^T(\Sigma_{g'} + \text{diag}(\Lambda))\boldsymbol{t}}{2}\right)$$

The PRODUCT part of the score is given by

$$\text{PRODUCT} = \mathbb{E}\left[it_1 z_1\right] \mathbb{E}\left[it_2 z_2\right] \exp\left(-\frac{\boldsymbol{t}^T(\text{diag}(\Sigma_{g'}) + \Lambda)\boldsymbol{t}}{2}\right)$$

It is not hard to see that JOINT equals PRODUCT. The same argument generalizes to $k > 2$. We next provide proof for the harder direction here. Suppose that $\Delta_F(\boldsymbol{t}|P) = 0$, i.e,

$$\mathbb{E}\left[\exp(i\boldsymbol{t}^T F\boldsymbol{x})\right] \exp\left(-\frac{\boldsymbol{t}^T \text{diag}\left(FSF^T\right)\boldsymbol{t}}{2}\right) - \prod_{j=1}^{k} \mathbb{E}\left[\exp(it_j(F\boldsymbol{x})_j)\right] \exp\left(-\frac{\boldsymbol{t}^T FSF^T\boldsymbol{t}}{2}\right) = 0 \tag{A.9}$$

We then prove that F must be of the form DB^{-1} for D being a permutation of a diagonal matrix. Then, taking the logarithm, we have

$$\begin{aligned}\ln\left(\mathbb{E}\left[\exp(i\boldsymbol{t}^T F\boldsymbol{x})\right]\right) - \sum_{j=1}^{k} \ln\left(\mathbb{E}\left[\exp(it_j(F\boldsymbol{x})_j)\right]\right) &= \frac{1}{2}\boldsymbol{t}^T\left(\text{diag}\left(FSF^T\right) - FSF^T\right)\boldsymbol{t} \\ &= \frac{1}{2}\boldsymbol{t}^T\left(\text{diag}\left(F\Sigma F^T\right) - F\Sigma F^T\right)\boldsymbol{t}\end{aligned}$$

Under the ICA model we have, $\boldsymbol{x} = B\boldsymbol{z} + \boldsymbol{g}$. Therefore, from Eq A.9, using the definition of the characteristic function of a Gaussian random variable and noting that $\text{cov}\,(\boldsymbol{g}) = \Sigma$, we have

$$\begin{aligned}\ln\left(\mathbb{E}\left[\exp(i\boldsymbol{t}^T F\boldsymbol{x})\right]\right) &= \ln\left(\mathbb{E}\left[\exp(i\boldsymbol{t}^T FB\boldsymbol{z})\right]\right) + \ln\left(\mathbb{E}\left[\exp(i\boldsymbol{t}^T F\boldsymbol{g})\right]\right) \\ &= \ln\left(\mathbb{E}\left[\exp(i\boldsymbol{t}^T FB\boldsymbol{z})\right]\right) - \frac{1}{2}\boldsymbol{t}^T F\Sigma F^T\boldsymbol{t}\end{aligned}$$

and similarly,

$$\sum_{j=1}^{k} \ln\left(\mathbb{E}\left[\exp(it_j(F\boldsymbol{x})_j)\right]\right) = \sum_{j=1}^{k} \ln\left(\mathbb{E}\left[\exp(it_j(FB\boldsymbol{z})_j)\right]\right) - \frac{1}{2}\boldsymbol{t}^T \text{diag}\left(F\Sigma F^T\right)\boldsymbol{t}$$

Therefore, from Eq A.9,

$$g_F(\boldsymbol{t}) := \ln\left(\mathbb{E}\left[\exp(i\boldsymbol{t}^T FB\boldsymbol{z})\right]\right) - \sum_{j=1}^{k} \ln\left(\mathbb{E}\left[\exp(it_j(FB\boldsymbol{z})_j)\right]\right)$$

must be a quadratic function of $\boldsymbol{t}$. It is important to note that the second term is an additive function w.r.t $t_1, t_2, \cdots t_k$. For simplicity, consider the case of $k = 2$ and assume that the joint and marginal characteristic functions of all signals are twice differentiable. Consider $\frac{\partial^2\left(g_F(\boldsymbol{t})\right)}{\partial t_1 \partial t_2}$, then

$$\sum_{k=1}^{2} \frac{\partial^2\left(g_F(\boldsymbol{t})\right)}{\partial t_1 \partial t_2} \equiv const. \tag{A.10}$$

To simplify the notation, let $\psi_j(t) := \mathbb{E}\left[e^{itz_j}\right]$, $f_j \equiv \log\psi_j$, $M := FB$. The above is then equivalent to

$$\sum_{k=1}^{2} M_{1k}M_{2k} f_k''(M_{1k}t_1 + M_{2k}t_2) \equiv const. \tag{A.11}$$

Note that M is invertible since F, B are invertible. Therefore, let $\mathbf{s} \equiv M^T\mathbf{t}$, then

$$\sum_{k=1}^{2} M_{1k}M_{2k} f_k''(s_k) \equiv const.$$

which implies that either $M_{1k}M_{2k} = 0$ or $f_k''(s_k) \equiv const$, for $k = 1, 2$. For any $k \in \{1, 2\}$, since z_k is non-Gaussian, its logarithmic characteristic function, i.e. f_k cannot be a quadratic function, so

$$M_{1k}M_{2k} = 0$$

which implies that each column of M has a zero. Since M is invertible, thus M is either a diagonal matrix or a permutation of a diagonal matrix. Now to consider $k > 2$, we just need to apply the above $k = 2$ argument to pairwise entries of $\{t_1, \cdots, t_k\}$ with other entries fixed. Hence proved. □

Proof of Theorem 3. We start with considering the following constrained optimization problem -

$$\sup_{\boldsymbol{u}^T C^{-1}\boldsymbol{u}=1} f\left(C^{-1}\boldsymbol{u}|P\right)$$

By an application of lagrange multipliers, for the optima $\boldsymbol{u}_{\text{opt}}$, we have

$$\nu C^{-1}\boldsymbol{u}_{opt} = C^{-1}\nabla f\left((C^{-1})^T\boldsymbol{u}_{\text{opt}}|P\right), \quad \boldsymbol{u}_{\text{opt}}^T C^{-1}\boldsymbol{u}_{\text{opt}} = 1$$

with multiplier $\nu \in \mathbb{R}$. This leads to a fixed-point iteration very similar to the one in [49], which is the motivation for considering such an optimization framework. We now introduce some notations to simplify the problem.

Let $\boldsymbol{u} = B\boldsymbol{\alpha}$, $z_i' := \frac{z_i}{d_i}$ and $a_i' := \text{var}(z_i') = \frac{a_i}{d_i^2}$, where $a_i := \text{var}(z_i)$. Then,

$$\sup_{\boldsymbol{u}^T C^{-1}\boldsymbol{u}=1} f\left(C^{-1}\boldsymbol{u}|P\right) = \sup_{\boldsymbol{\alpha}^T D^{-1}\boldsymbol{\alpha}=1} f\left(D^{-1}\boldsymbol{\alpha}|\boldsymbol{z}\right)$$

We will use $f(\boldsymbol{u}|P)$ or $f(\boldsymbol{u}|\boldsymbol{x})$ where $\boldsymbol{x} \sim P$ interchangeably. By Assumption 1-(a), we see that $f\left(D^{-1}\boldsymbol{\alpha}|\boldsymbol{z}\right) = \sum_{i=1}^{k} f\left(\frac{\alpha_i}{d_i}|z_i\right)$. For simplicity of notation, we will use $h_i(\alpha_i/d_i) = f\left(\alpha_i/d_i|z_i\right)$, since the functional form can be different for different random variables z_i. Therefore, we seek to characterize the fixed points of the following optimization problem -

$$\begin{aligned} &\sup_{\boldsymbol{\alpha}} \sum_{i=1}^{k} h_i(\alpha_i/d_i) \\ &s.t \quad \sum_{i=1}^{k} \frac{\alpha_i^2}{d_i} = 1, \\ &\quad d_i \neq 0 \ \ \forall\, i \in [k] \end{aligned} \tag{A.12}$$

 https://doi.org/10.52202/079017-4217

where $\boldsymbol{\alpha}, \boldsymbol{d} \in \mathbb{R}^k$. We have essentially moved from optimizing in the $\langle ., . \rangle_{C^{-1}}$ space to the $\langle ., . \rangle_{D^{-1}}$ space. We find stationary points of the Lagrangian

$$\mathcal{L}(\boldsymbol{\alpha}, \lambda) := \sum_{i=1}^{k} h_i \left(\frac{\alpha_i}{d_i} \right) - \lambda \left(\sum_{i=1}^{k} \frac{\alpha_i^2}{d_i} - 1 \right) \tag{A.13}$$

The components of the gradient of $\mathcal{L}(\boldsymbol{\alpha}, \lambda)$ w.r.t $\boldsymbol{\alpha}$ are given as

$$\frac{\partial}{\partial \alpha_j} \mathcal{L}(\boldsymbol{\alpha}, \lambda) = \frac{1}{d_j} h_j' \left(\frac{\alpha_j}{d_j} \right) - \lambda \frac{\alpha_j}{d_j},$$

At the fixed point $(\boldsymbol{\alpha}, \lambda)$ we have,

$$\forall j \in [k], h_j' \left(\frac{\alpha_j}{d_j} \right) - \lambda \alpha_j = 0$$

By Assumption 1-(b), for $\alpha_i = 0$, the above is automatically zero. However, for $\{j : \alpha_j \neq 0\}$, we have,

$$\lambda = \frac{h_j' \left(\frac{\alpha_j}{d_j} \right)}{\alpha_j} \tag{A.14}$$

So, for all $\{j : \alpha_j \neq 0\}$, we must have the same value and sign of $\frac{1}{\alpha_j} h_j' \left(\frac{\alpha_j}{d_j} \right)$. By assumption in the theorem statement, $h_i'''(x)$ does not change sign on the half-lines $x > 0$ and $x < 0$. Along with Assumption 1-(b), this implies that

$$\forall\, x \in [0, \infty),\ \text{sgn}(h_i(x)) = \text{sgn}(h_i'(x)) = \text{sgn}(h_i''(x)) = \text{sgn}(h_i'''(x)) = \kappa_1 \tag{A.15}$$

$$\forall\, x \in (-\infty, 0],\ \text{sgn}(h_i(x)) = -\text{sgn}(h_i'(x)) = \text{sgn}(h_i''(x)) = -\text{sgn}(h_i'''(x)) = \kappa_2 \tag{A.16}$$

where $\kappa_1, \kappa_2 \in \{-1, 1\}$ are constants. Furthermore, we note that Assumption 1-(d) ensures that $h_i(x)$ is a symmetric function. Therefore, $\forall\, x \in \mathbb{R}, \text{sgn}(h_i(x)) = \text{sgn}(h_i''(x)) = \kappa, \kappa \in \{-1, 1\}$. Then, since $d_i = h_i''(u_i)$,

$$\text{sgn}\left(d_i\right) = \text{sgn}\left(h_i(x)\right), \forall x \in \mathbb{R} \tag{A.17}$$

Now, using A.14,

$$\begin{aligned}
\text{sgn}(\lambda) &= \text{sgn}\left(h_j' \left(\frac{\alpha_j}{d_j} \right) \right) \times \text{sgn}(\alpha_j) \\
&= \text{sgn}\left(\frac{\alpha_j}{d_j} \right) \times \text{sgn}\left(h_j \left(\frac{\alpha_j}{d_j} \right) \right) \times \text{sgn}(\alpha_j), \text{using Eq } A.15 \text{ and } A.16 \\
&= \text{sgn}\left(d_j\right) \times \text{sgn}\left(h_j \left(\frac{\alpha_j}{d_j} \right) \right) \\
&= 1, \text{using Eq } A.17
\end{aligned} \tag{A.18}$$

Keeping this in mind, we now compute the Hessian, $H \in \mathbb{R}^{k \times k}$, of the lagrangian, $\mathcal{L}(\boldsymbol{\alpha}, \lambda)$ at the fixed point, $(\boldsymbol{\alpha}, \lambda)$. Recall that, we have $h_i'(0) = 0$ and $h_i''(0) = 0$. Thus, for $\{i : \alpha_i = 0\}$,

$$H_{ii} = -\lambda / d_i \tag{A.19}$$

This implies that $\text{sgn}\left(d_i H_{ii}\right) = \text{sgn}\left(\lambda\right) = -1$ for $\{i : \alpha_i = 0\}$, using Eq A.18.

For $\{i : \alpha_i \neq 0\}$, we have

$$\begin{aligned}
H_{ij} = \frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \mathcal{L}(\boldsymbol{\alpha}, \lambda) \Big|_{\boldsymbol{\alpha}} &= \mathbb{1}(i = j) \left[\frac{h_i'' \left(\frac{\alpha_i}{d_i} \right)}{d_i^2} - \frac{\lambda}{d_i} \right] \\
&= \mathbb{1}(i = j) \left[\frac{h_i'' \left(\frac{\alpha_i}{d_i} \right)}{d_i^2} - \frac{h_i' \left(\frac{\alpha_i}{d_i} \right)}{\alpha_i d_i} \right], \qquad \text{for } \alpha_i \neq 0, \text{ using } A.14 \\
&= \mathbb{1}(i = j) \frac{1}{d_i \alpha_i} \left[\frac{\alpha_i}{d_i} h_i'' \left(\frac{\alpha_i}{d_i} \right) - h_i' \left(\frac{\alpha_i}{d_i} \right) \right], \qquad \text{for } \alpha_i \neq 0
\end{aligned}$$

We consider the pseudo inner product space $\langle ., . \rangle_{D^{-1}}$ for optimizing $\boldsymbol{\alpha}$. Furthermore, since we are in this pseudo-inner product space, we have

$$\langle \boldsymbol{v}, H\boldsymbol{v} \rangle_{D^{-1}} = \boldsymbol{v} D^{-1} H \boldsymbol{v}$$

So we will consider the positive definite-ness of the matrix $\tilde{H} := D^{-1}H$ to characterize the fixed points. Recall that for a differentiable convex function f, we have $\forall\ x, y \in dom(f) \subseteq \mathbb{R}^n$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

Therefore, for $\{i : \alpha_i \neq 0\}$, we have

$$\begin{aligned}
\text{sgn}\left(d_i H_{ii}\right) &= \text{sgn}\left(d_i\right) \times \text{sgn}\left(d_i \alpha_i\right) \times \text{sgn}\left(\frac{\alpha_i}{d_i} h_i''\left(\frac{\alpha_i}{d_i}\right) - h_i'\left(\frac{\alpha_i}{d_i}\right)\right) \\
&= \text{sgn}\left(\alpha_i\right) \times \text{sgn}\left(h_i'''\left(\frac{\alpha_i}{d_i}\right)\right), \text{using convexity/concavity of } h'(.) \\
&= \text{sgn}\left(\alpha_i\right) \times \text{sgn}\left(h_i'\left(\frac{\alpha_i}{d_i}\right)\right), \text{using } A.15 \text{ and } A.16 \\
&= \text{sgn}\left(\alpha_i\right) \times \text{sgn}\left(\lambda\right) \times \text{sgn}\left(\alpha_i\right), \text{using } A.14 \\
&= \text{sgn}\left(\lambda\right) \\
&= 1, \text{using Eq } A.17
\end{aligned} \tag{A.20}$$

Let $S := \{i : \alpha_i \neq 0\}$. We then have the following cases -

1. Assume that only one α_i is nonzero, then $\forall \boldsymbol{v}$ orthogonal to this direction, $\langle \boldsymbol{v}, H\boldsymbol{v} \rangle_{D^{-1}} < 0$ by Eq A.19. Thus this gives a local maxima.
2. Assume more than one α_i are nonzero, but $|S| < k$. Then we have for $i \notin S$, $\tilde{H}_{jj} < 0$ using A.19. For $i \in S$, $\tilde{H}_{ii} > 0$ from Eq A.20. Hence these are saddle points.
3. Assume we have $S = [k]$, i.e. $\forall i$, $\alpha_i \neq 0$. In this case, $\tilde{H} \succ 0$. So, we have a local minima.

This completes our proof.

□

Theorem A.1. *Consider the data generated from the noisy ICA model (Eq 1). If $f(\boldsymbol{u}|P)$ is defined as in Eq 4 or Eq 5, we have $\nabla^2 f(\boldsymbol{u}|P) = BD_u B^T$, for some diagonal matrix D_u, which can be different for the differerent contrast functions.*

Remark 4. *The above theorem is useful because it shows that the Hessian of the contrast functions based on the CHF (eq 4 or CGF (eq 5) is of the form BDB^T, where D is some diagonal matrix. This matrix can be precomputed at some vector $\boldsymbol{u}$ and used as the matrix C in the power iteration update (see eq 3) in Algorithm 2 of [49].*

Proof of Theorem A.1. First, we show that the CHF-based contrast function (see Eq 4), the Hessian, is of the correct form.

CHF-based contrast function

$$f(\boldsymbol{u}|P) = \log \mathbb{E} \exp(i\boldsymbol{u}^T B\boldsymbol{z}) + \log \mathbb{E} \exp(-i\boldsymbol{u}^T B\boldsymbol{z}) + \boldsymbol{u}^T B \,\text{diag}(\boldsymbol{a}) B^T \boldsymbol{u},$$

where $\text{diag}(\boldsymbol{a})$ is the covariance matrix of $\boldsymbol{z}$. For simplicity of notation, define $\phi(\boldsymbol{z}; \boldsymbol{u}) := \mathbb{E}\left[\exp(i\boldsymbol{u}^T B\boldsymbol{z})\right]$.

$$\nabla f(\boldsymbol{u}|P) = iB \left(\underbrace{\frac{\mathbb{E}\left[\exp(i\boldsymbol{u}^T B\boldsymbol{z})\boldsymbol{z}\right]}{\phi(\boldsymbol{z}; \boldsymbol{u})}}_{\mu(\boldsymbol{u}; \boldsymbol{z})} - i \frac{\mathbb{E}\left[\exp(-i\boldsymbol{u}^T \boldsymbol{z})\boldsymbol{z}\right]}{\phi(\boldsymbol{z}; -\boldsymbol{u})} \right) + 2B \,\text{diag}(\boldsymbol{a}) B^T \boldsymbol{u}$$

 https://doi.org/10.52202/079017-4217

Define

$$H(\boldsymbol{u};\boldsymbol{z}) := \frac{\mathbb{E}\left[\exp(i\boldsymbol{u}^T B\boldsymbol{z})\boldsymbol{z}\boldsymbol{z}^T\right]}{\phi(\boldsymbol{z};\boldsymbol{u})} - \mu(\boldsymbol{u};\boldsymbol{z})\mu(\boldsymbol{u};\boldsymbol{z})^T.$$

Taking a second derivative, we have:

$$\nabla^2 f(\boldsymbol{u}|P) = B\left(-H(\boldsymbol{u};\boldsymbol{z}) - H(\boldsymbol{z};-\boldsymbol{u}) + 2\,\mathrm{diag}(\boldsymbol{a})\right)B^T$$

All we have to show at this point is that $H(\boldsymbol{u};\boldsymbol{z})$ is diagonal. We will evaluate the k,ℓ entry. Let B_i denote the i^{th} column of B. Let $Y_{\setminus k,\ell} := \boldsymbol{u}^T B \sum_{j\neq k,\ell} z_j$. Using independence of the components of $\boldsymbol{z}$, we have, for $k \neq \ell$,

$$\frac{\mathbb{E}\left[\exp(i\boldsymbol{u}^T B_k z_k + i\boldsymbol{u}^T B_\ell z_\ell + Y_{\setminus k\ell})z_k z_\ell\right]}{\phi(\boldsymbol{z};\boldsymbol{u})} = \frac{\mathbb{E}\left[\exp(i\boldsymbol{u}^T B_k z_k)z_k\right]\mathbb{E}\left[\exp(i\boldsymbol{u}^T B_\ell z_\ell)z_\ell\right]}{\mathbb{E}\left[\exp(i\boldsymbol{u}^T B_k z_k)\right]\mathbb{E}\left[\exp(i\boldsymbol{u}^T B_\ell z_\ell)\right]} \tag{A.21}$$

Now we evaluate:

$$\boldsymbol{e}_k^T\mu(\boldsymbol{u};\boldsymbol{z}) := \frac{\mathbb{E}\left[\exp(i\boldsymbol{u}^T B\boldsymbol{z})z_k\right]}{\phi(\boldsymbol{z};\boldsymbol{u})} = \frac{\mathbb{E}\left[\exp(i\boldsymbol{u}^T B_k z_k)z_k\right]}{\mathbb{E}\left[\exp(i\boldsymbol{u}^T B_k z_k)\right]} \tag{A.22}$$

Using Eqs A.21 and A.22, we see that indeed $H(\boldsymbol{u};\boldsymbol{z})$ is diagonal.

Now, we prove the statement about the CGF-based contrast function.

CGF-based contrast function

We have,

$$\nabla f\left(\boldsymbol{u}|P\right) = \frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)B\boldsymbol{z}\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]} - B\,\mathrm{diag}\left(\boldsymbol{a}\right)B^T\boldsymbol{u},$$

$$\nabla^2 f\left(\boldsymbol{u}|P\right) = \frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)B\boldsymbol{z}\boldsymbol{z}^T B^T\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]} - \frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)B\boldsymbol{z}\right]\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\boldsymbol{z}^T B^T\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]^2} - B\,\mathrm{diag}\left(\boldsymbol{a}\right)B^T$$

$$= B\left[\underbrace{\frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\boldsymbol{z}\boldsymbol{z}^T\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]} - \frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\boldsymbol{z}\right]\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\boldsymbol{z}^T\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]^2}}_{\text{Covariance of } \boldsymbol{z}\sim\mathbb{P}(\boldsymbol{z}).\,\frac{\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]}} - \mathrm{diag}\left(\boldsymbol{a}\right)\right]B^T$$

The new probability density $\mathbb{P}\left(\boldsymbol{z}\right)\cdot\frac{\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)}{\mathbb{E}[\exp(\boldsymbol{u}^T B\boldsymbol{z})]}$ is an exponential tilt of the original pdf, and since $\{z_i\}_{i=1}^d$ are independent, and the new tilted density also factorizes over the z_i's, therefore, the covariance under this tilted density is also diagonal. □

A.1.1 Uniform convergence

In this section, we provide the proof for Theorem 2. First, we state some preliminary results about the uniform convergence of smooth function classes which will be useful for our proofs.

A.1.1.1 Preliminaries

Let $D := \sup_{\theta,\tilde{\theta}\in\mathbb{T}}\rho_X\left(\theta,\tilde{\theta}\right)$ denote the diameter of set $\mathbb{T}$, and let $\mathcal{N}_X\left(\delta;\mathbb{T}\right)$ denote the δ-covering number of $\mathbb{T}$ in the ρ_X metric. Then, we have the following standard result:

Proposition A.2. *(Proposition 5.17 from [51]) Let $\{X_\theta, \theta\in\mathbb{T}\}$ be a zero-mean subgaussian process with respect to the metric ρ_X. Then for any $\delta\in[0,D]$ such that $\mathcal{N}_X\left(\delta;\mathbb{T}\right)\geq 10$, we have*

$$\mathbb{E}\left[\sup_{\theta,\tilde{\theta}\in\mathbb{T}}\left(X_\theta - X_{\tilde{\theta}}\right)\right] \leq 2\mathbb{E}\left[\sup_{\gamma,\gamma'\in\mathbb{T};\rho_X(\gamma,\gamma')\leq\delta}\left(X_\theta - X_{\tilde{\theta}}\right)\right] + 4\sqrt{D^2\log\left(\mathcal{N}_X\left(\delta;\mathbb{T}\right)\right)}$$

Remark 5. *For zero-mean subgaussian processes, Proposition A.2 implies,* $\mathbb{E}\left[\sup_{\theta\in\mathbb{T}} X_\theta\right] \leq \mathbb{E}\left[\sup_{\theta,\tilde{\theta}\in\mathbb{T}}\left(X_\theta - X_{\tilde{\theta}}\right)\right]$

Proposition A.3. *(Theorem 4.10 from [51]) For any b-uniformly bounded class of functions* $\mathcal{C}$*, any positive integer* $n \geq 1$*, any scalar* $\delta \geq 0$*, and a set of i.i.d datapoints* $\left\{X^{(i)}\right\}_{i\in[n]}$ *we have*

$$\sup_{f\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} f\left(X^{(i)}\right) - \mathbb{E}\left[f\left(X\right)\right]\right| \leq 2\frac{1}{n}\mathbb{E}_{X,\epsilon}\left[\sup_{f\in\mathcal{C}}\left|\sum_{i=1}^{n}\epsilon_i f\left(X^{(i)}\right)\right|\right] + \delta$$

with probability at least $1 - \exp\left(-\frac{n\delta^2}{2b^2}\right)$*. Here* $\{\epsilon_i\}_{i\in[n]}$ *are i.i.d rademacher random variables.*

A.1.1.2 Proof of theorem 2

For dataset $\left\{\boldsymbol{x}^{(i)}\right\}_{i\in[n]}, \boldsymbol{x}^{(i)} \in \mathbb{R}^k$, consider the following definitions:

$$\phi(\boldsymbol{t}, F|P) := \mathbb{E}_{\boldsymbol{x}}\left[\exp(i\boldsymbol{t}^T F\boldsymbol{x})\right], \quad \phi(\boldsymbol{t}, F|\hat{P}) := \frac{1}{n}\sum_{j=1}^{n}\exp(i\boldsymbol{t}^T F\boldsymbol{x}^{(j)}) \tag{A.23}$$

$$\psi(\boldsymbol{t}, F|P) := \prod_{j=1}^{k}\mathbb{E}_{\boldsymbol{x}}\left[\exp(it_j(F\boldsymbol{x})_j)\right], \quad \psi(\boldsymbol{t}, F|\hat{P}) := \frac{1}{n}\prod_{j=1}^{k}\sum_{r=1}^{n}\exp(it_j(F\boldsymbol{x}^{(r)})_j) \tag{A.24}$$

Let $\Delta(\boldsymbol{t}, F|\hat{P})$ be the empirical version where the expectation is replaced by sample averages. Now define:

$$\Delta(\boldsymbol{t}, F|P) = \left|\phi(\boldsymbol{t}, F|P)\exp\left(-\boldsymbol{t}^T\operatorname{diag}(FSF^T)\boldsymbol{t}\right) - \psi(\boldsymbol{t}, F|P)\exp\left(-\boldsymbol{t}^T FSF^T\boldsymbol{t}\right)\right|$$
$$\Delta(\boldsymbol{t}, F|\hat{P}) = \left|\phi(\boldsymbol{t}, F|\hat{P})\exp\left(-\boldsymbol{t}^T\operatorname{diag}(F\widehat{S}F^T)\boldsymbol{t}\right) - \psi(\boldsymbol{t}, F|\hat{P})\exp\left(-\boldsymbol{t}^T F\widehat{S}F^T\boldsymbol{t}\right)\right|$$

Theorem A.4. *Let* $\mathcal{F} := \{F \in \mathbb{R}^{k\times k} : \|F\| \leq 1\}$*. Assume that* $\boldsymbol{x} \sim$ *subgaussian*(σ)*. We have:*

$$\sup_{F\in\mathcal{F}}|\mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\Delta(\boldsymbol{t}, F|P) - \mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\Delta(\boldsymbol{t}, F|\hat{P})| = O_P\left(\sqrt{\frac{k^2\|S\|\max(k, \sigma^4\|S\|)\log^2(nC_k)}{n}}\right)$$

where $C_k := \max(1, k\log(n)\operatorname{Tr}(S))$.

Proof. Define $\mathcal{E} = \{\boldsymbol{t} : \|\boldsymbol{t}\| \leq \sqrt{2k\log n}\}$. We will use the fact that $P(\mathcal{E}^c) \leq 2/n$ (Example 2.12, [51]). Also note that by construction, $\Delta(\boldsymbol{t}, F|P) \leq 1$. Observe that:

$$\begin{aligned}
&\sup_{F\in\mathcal{F}}|\mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\Delta(\boldsymbol{t}, F|P) - \mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\Delta(\boldsymbol{t}, F|\hat{P})| \\
&\leq \sup_{F\in\mathcal{F}}\mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}|\Delta(\boldsymbol{t}, F|P) - \Delta(\boldsymbol{t}, F|\hat{P})| \\
&\leq \mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\sup_{F\in\mathcal{F}}|\Delta(\boldsymbol{t}, F|P) - \Delta(\boldsymbol{t}, F|\hat{P})| \\
&\leq \mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\left[\sup_{F\in\mathcal{F}}\left|\Delta(\boldsymbol{t}, F|P) - \Delta(\boldsymbol{t}, F|\hat{P})\right|\middle|\mathcal{E}\right] + \mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\left[\sup_{F\in\mathcal{F}}|\Delta(\boldsymbol{t}, F|P) - \Delta(\boldsymbol{t}, F|\hat{P})\middle||\mathcal{E}^c\right]P(\mathcal{E}^c) \\
&\leq \mathbb{E}_{\boldsymbol{t}\in N(0,I_k)}\left[\sup_{F\in\mathcal{F}}\left|\Delta(\boldsymbol{t}, F|P) - \Delta(\boldsymbol{t}, F|\hat{P})\right|\middle|\mathcal{E}\right] + 2/n \\
&= O_P\left(\sqrt{\frac{k^2\|S\|\max(k, \sigma^4\|S\|)\log^2(nC_k)}{n}}\right) + 2/n, \text{ using Theorem } A.5
\end{aligned}$$

The second inequality follows from Jensen's inequality, the fourth follows from $\mathbb{E}[\sup(.)] \leq \sup(\mathbb{E}[.])$. □

 https://doi.org/10.52202/079017-4217

Theorem A.5. *Let $\mathcal{F} := \{F \in \mathbb{R}^{k\times k} : \|F\| \leq 1\}$. Assume that $\boldsymbol{x} \sim$ subgaussian(σ). We have:*

$$\sup_{F\in\mathcal{F}} |\Delta(\boldsymbol{t}, F|P) - \Delta(\boldsymbol{t}, F|\hat{P})| = O_P\left(\|\boldsymbol{t}\|\sqrt{\frac{k\|S\|\max(k, \sigma^4\|S\|)\log(nC_t)}{n}}\right)$$

where $C_t := \max(1, \|\boldsymbol{t}\|^2 \mathrm{Tr}(S))$.

Proof.

$$\begin{aligned}
\Delta(\boldsymbol{t}, F|P) - \Delta(\boldsymbol{t}, F|\hat{P}) \leq &\left|\phi(\boldsymbol{t}, F|P) - \phi(\boldsymbol{t}, F|\hat{P})\right| \exp\left(-\boldsymbol{t}^T \operatorname{diag}(FSF^T)\boldsymbol{t}\right) \\
&+ \phi(\boldsymbol{t}, F|\hat{P}) \left|\exp\left(-\boldsymbol{t}^T \operatorname{diag}(F\widehat{S}F^T)\boldsymbol{t}\right) - \exp\left(-\boldsymbol{t}^T \operatorname{diag}(FSF^T)\boldsymbol{t}\right)\right| \\
&+ \left|\psi(\boldsymbol{t}, F|P) - \psi(\boldsymbol{t}, F|\hat{P})\right| \exp\left(-\boldsymbol{t}^T FSF^T\boldsymbol{t}\right) \\
&+ \psi(\boldsymbol{t}, F|\hat{P}) \left|\exp\left(-\boldsymbol{t}^T F\widehat{S}F^T\boldsymbol{t}\right) - \exp\left(-\boldsymbol{t}^T FSF^T\boldsymbol{t}\right)\right|
\end{aligned}$$

Finally, for some $S' = \lambda S + (1-\lambda)\hat{S}$,

$$\begin{aligned}
|\exp(-\boldsymbol{t}^T FSF^T\boldsymbol{t}) - \exp(-\boldsymbol{t}^T F\hat{S}F^T\boldsymbol{t})| &= \left|\left\langle \partial_S \exp(-\boldsymbol{t}^T FSF^T\boldsymbol{t})\big|_{S'}, \hat{S} - S\right\rangle\right| \\
&\leq \exp(-\boldsymbol{t}^T FS'F^T\boldsymbol{t}) \left|\boldsymbol{t}^T F(S-\hat{S})F^T\boldsymbol{t}\right| \\
&\leq \|\boldsymbol{t}\|^2\|S-\hat{S}\| = \|\boldsymbol{t}\|^2 O_P\left(\sigma^2\|S\|\sqrt{\frac{k}{n}}\right)
\end{aligned}$$

where the last result follows from Theorem 4.7.1 in [47]. Now, for the second term, observe that:

$$\begin{aligned}
&\left|\phi(\boldsymbol{t}, F|\hat{P}) \left(\exp\left(-\boldsymbol{t}^T \operatorname{diag}(F\widehat{S}F^T)\boldsymbol{t}\right) - \exp\left(-\boldsymbol{t}^T \operatorname{diag}(FSF^T)\boldsymbol{t}\right)\right)\right| \\
&\leq \exp\left(-\boldsymbol{t}^T \operatorname{diag}(FSF^T)\boldsymbol{t}\right) \left|\exp\left(-\boldsymbol{t}^T \operatorname{diag}(F(\widehat{S}-S)F^T)\boldsymbol{t}\right) - 1\right| \qquad \text{(A.25)}
\end{aligned}$$

We next have that, with probability at least $1 - 1/n$,

$$\left\|\operatorname{diag}\left(F(S-\widehat{S})F^T\right)\right\|_2 \leq \left\|F(S-\widehat{S})F^T\right\|_2 \leq C\left(\sigma^2\|S\|\sqrt{\frac{k\log n}{n}}\right)$$

Thus with probability at least $1 - 1/n$, using the inequality $|1 - e^x| \leq 2|x|$, $\forall x \in [-1, 1]$, Eq A.25 leads to:

$$\left|\phi(\boldsymbol{t}, F|\hat{P}) \left(\exp\left(-\boldsymbol{t}^T \operatorname{diag}(F\widehat{S}F^T)\boldsymbol{t}\right) - \exp\left(-\boldsymbol{t}^T \operatorname{diag}(FSF^T)\boldsymbol{t}\right)\right)\right| \leq 2C\|\boldsymbol{t}\|^2\left(\sigma^2\|S\|\sqrt{\frac{k\log n}{n}}\right)$$

Note that the matrices $\operatorname{diag}\left(F(S-\widehat{S})F^T\right)$ and FSF^T are positive semi-definite and $\boldsymbol{t}$ is unit-norm. Observe that, using Lemmas A.6 and A.7, we have:

$$\begin{aligned}
&\sup_{F\in\mathcal{F}} |\Delta(\boldsymbol{t}, F|P) - \Delta(\boldsymbol{t}, F|\hat{P})| \\
&\leq O_P\left(\|\boldsymbol{t}\|\sqrt{\frac{k\,\mathrm{Tr}(S)\log(nC_t)}{n}}\right) + O_P\left(\|\boldsymbol{t}\|\sigma^2\|S\|\sqrt{\frac{k\log n}{n}}\right) + O_P\left(\|\boldsymbol{t}\|\sqrt{\frac{k^2\|S\|\log n}{n}}\right) \\
&= O_P\left(\|\boldsymbol{t}\|\sqrt{\frac{k\|S\|\max(k, \sigma^4\|S\|)\log(nC_t)}{n}}\right)
\end{aligned}$$

□

Lemma A.6. *Define* $\mathcal{F} := \{F \in \mathbb{R}^{k\times k} : \|F\| \leq 1\}$. *Let* $\left\{\boldsymbol{x}^{(i)}\right\}_{i\in[n]}$ *be i.i.d samples from a subgaussian*(σ) *distribution. We have:*

$$\sup_{F\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\exp(i\boldsymbol{t}^TF\boldsymbol{x}^{(j)}) - \mathbb{E}_{\boldsymbol{x}}\left[\exp(i\boldsymbol{t}^TF\boldsymbol{x})\right]\right| = O_P\left(\|\boldsymbol{t}\|\sqrt{\frac{k\,\mathrm{Tr}(S)\log(nC_t)}{n}}\right)$$

where $C_t := \max(1, \|\boldsymbol{t}\|^2\mathrm{Tr}(S))$.

Proof. Let $\phi(\boldsymbol{t}, F; \boldsymbol{x}^{(i)}) = \exp(i\boldsymbol{t}^TF\boldsymbol{x}^{(i)})$. Next, we note that $\|F\|_2 \leq 1$ imply $\left\|F^T\boldsymbol{t}\right\| \leq \|\boldsymbol{t}\|$. Therefore,

$$\sup_{F\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\exp(i\boldsymbol{t}^TF\boldsymbol{x}^{(j)}) - \mathbb{E}_{\boldsymbol{x}}\left[\exp(i\boldsymbol{t}^TF\boldsymbol{x})\right]\right| \leq \sup_{\boldsymbol{u}\in\mathbb{R}^k,\|\boldsymbol{u}\|\leq 1}\left|\frac{1}{n}\sum_{j=1}^{n}\exp(i\boldsymbol{u}^T\boldsymbol{x}^{(j)}) - \mathbb{E}_{\boldsymbol{x}}\left[\exp(i\boldsymbol{u}^T\boldsymbol{x})\right]\right|$$

Let $f_X(\boldsymbol{u}) = \sum_i \epsilon_i\xi(\boldsymbol{u}; \boldsymbol{x}^{(i)})$. We will argue for unit vectors $\boldsymbol{u}$, and later scale them by $\|\boldsymbol{t}\|$. Define $\mathcal{B}_k := \left\{\boldsymbol{u} \in \mathbb{R}^k, \|\boldsymbol{u}\| \leq 1\right\}$, $\xi(\boldsymbol{u}; \boldsymbol{x}) := \exp(i\boldsymbol{u}^T\boldsymbol{x})$ and let

$$d(\boldsymbol{u}, \boldsymbol{u}')^2 := \sum_i (\xi(\boldsymbol{u}; \boldsymbol{x}^{(i)}) - \xi(\boldsymbol{u}'; \boldsymbol{x}^{(i)}))^2$$

Then we have,

$$\|\nabla_{\boldsymbol{u}}\exp(i\boldsymbol{u}^T\boldsymbol{x})\|_2 = \|\boldsymbol{x}\||-\sin(\boldsymbol{u}^T\boldsymbol{x}) + i\cos(\boldsymbol{u}^T\boldsymbol{x})| \leq \|\boldsymbol{x}\|$$

Then,

$$\begin{aligned}\left|\xi(\boldsymbol{u}; \boldsymbol{x}^{(i)}) - \xi(\boldsymbol{t}; \boldsymbol{x}^{(i)})\right| &\leq \min\left(\|\nabla_{\boldsymbol{u}''}\xi(\boldsymbol{u}; \boldsymbol{x}^{(i)})\|_2\|\boldsymbol{u} - \boldsymbol{u}'\|, 2\right)\\ &\leq \min\left(\|\boldsymbol{x}^{(i)}\|_2\|\boldsymbol{u} - \boldsymbol{u}'\|, 2\right)\end{aligned}$$

Let $\widehat{S} := \sum_i \boldsymbol{x}^{(i)}\left(\boldsymbol{x}^{(i)}\right)^T/n$ and $\tau_n := n\,\mathrm{Tr}\left(\widehat{S}\right)$. We next have,

$$\begin{aligned}D^2 := \max_{\boldsymbol{u},\boldsymbol{u}'} d(\boldsymbol{u}, \boldsymbol{u}')^2 &\leq \min\left\{4n, \max_{\boldsymbol{u},\boldsymbol{u}'}\sum_i\|\boldsymbol{x}^{(i)}\|_2^2\|\boldsymbol{u} - \boldsymbol{u}'\|^2\right\}\\ &\leq 4\min\{n, \tau_n\}\end{aligned}$$

Next, we bound the covering number $N(\delta; \mathcal{B}_k, d_{\boldsymbol{u}})$. Note that $N(\delta; \mathcal{B}_k, d_{\boldsymbol{u}}) \leq N(\delta/\sqrt{\tau_n}; \mathcal{B}_k, \|.\|_2)$ since,

$$\begin{aligned}d(\boldsymbol{u}, \boldsymbol{u}')^2 &= \sum_i (\xi(\boldsymbol{u}; \boldsymbol{x}^{(i)}) - \xi(\boldsymbol{u}'; \boldsymbol{x}^{(i)}))^2\\ &\leq \sum_i\|\boldsymbol{x}^{(i)}\|_2^2\|\boldsymbol{u} - \boldsymbol{u}'\|^2\\ &= \sum_i \mathrm{Tr}\left(\boldsymbol{x}^{(i)}\left(\boldsymbol{x}^{(i)}\right)^T\right)\|\boldsymbol{u} - \boldsymbol{u}'\|^2\end{aligned}$$

Using Cauchy-Schwarz, we have:

$$|f_X(\boldsymbol{u}) - f_X(\boldsymbol{u}')| \leq \sum_{i=1}^{n}\epsilon_i|\xi(\boldsymbol{u}; \boldsymbol{x}^{(i)}) - \xi(\boldsymbol{u}; \boldsymbol{x}^{(i)})| \leq \sqrt{nd(\boldsymbol{u}, \boldsymbol{u}')^2}$$

 https://doi.org/10.52202/079017-4217

Therefore, we have $N(\delta; \mathcal{B}_k, d_{\boldsymbol{u}}) \leq \left(\frac{3\sqrt{\tau_n}}{\delta}\right)^k$. Therefore, using Proposition A.2,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x},\epsilon}\left[\sup_{F\in\mathcal{F}}|f_X(\boldsymbol{u}) - f_X(\boldsymbol{u}')|\right] &\leq 2\mathbb{E}_{\boldsymbol{x}}\left[\sup_{d(\boldsymbol{u},\boldsymbol{u}')\leq\delta}|f_X(\boldsymbol{u}) - f_X(\boldsymbol{u}')|\right] \;+\; 4\mathbb{E}_{\boldsymbol{x}}\left[D\sqrt{\log N(\delta;\mathcal{B}_k,d_{\boldsymbol{u}})}\right] \\
&\leq 2\mathbb{E}_{\boldsymbol{x}}\left[\inf_{\delta>0}\left\{\delta\sqrt{n} + 8\sqrt{\min\{n,\tau_n\}}\sqrt{2k\log\left(\frac{3\sqrt{\tau_n}}{\delta}\right)}\right\}\right] \\
&= O\left(\mathbb{E}_{\boldsymbol{x}}\sqrt{(k\min(n,\tau_n)\log(\max(\tau_n,n))}\right), \text{ setting } \delta := \sqrt{\min(1,\tau_n/n)} \\
&= \sqrt{kn}O\left(\sqrt{\min(1,\mathrm{Tr}(S))\log(n\max(1,\mathrm{Tr}(S)))}\right), \text{ Cauchy-Schwarz inequality}
\end{aligned}$$

Now we recall that $\|F^T\boldsymbol{t}\| \leq \|\boldsymbol{t}\|$, since all vectors $\boldsymbol{t}$ appear as $\boldsymbol{t}^T F^T \boldsymbol{x}^{(i)}$, this simply leads to scaling τ and $\|S\|$ by $\|\boldsymbol{t}\|^2$. Dividing both sides by n, we have

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{x},\epsilon}\left[\sup_{F\in\mathcal{F}}|f_X(\boldsymbol{t}) - f_X(\boldsymbol{t}')|\right] \leq \sqrt{\frac{k}{n}}O\left(\sqrt{\min(1,\|\boldsymbol{t}\|^2\mathrm{Tr}(S))\log(n\max(1,\|\boldsymbol{t}\|^2\mathrm{Tr}(S)))}\right)$$

Thus, using Proposition A.3 and using the definition of C_t, we have,

$$\sup_{F\in\mathcal{F}}|\phi(\boldsymbol{t},F|\hat{P}) - \phi(\boldsymbol{t},F|P)| = \sqrt{\frac{k}{n}}O\left(\sqrt{\|\boldsymbol{t}\|^2\mathrm{Tr}(S)\log(nC_t)}\right) = \|\boldsymbol{t}\|\sqrt{\frac{k}{n}}O\left(\sqrt{\mathrm{Tr}(S)\log(nC_t)}\right)$$

□

Lemma A.7. *Let $\mathcal{F} = \{F \in \mathbb{R}^{k\times k} : \|F\| \leq 1\}$. We have:*

$$\sup_{F\in\mathcal{F}}\left|\psi(\boldsymbol{t},F|\hat{P}) - \psi(\boldsymbol{t},F|P)\right| \leq O_P\left(\|\boldsymbol{t}\|\sqrt{\frac{k^2\|S\|\log n}{n}}\right)$$

Proof. Let $\mathcal{B}_k := \left\{\boldsymbol{u} \in \mathbb{R}^k, \|\boldsymbol{u}\| \leq 1\right\}$ Now define $\psi_j(\boldsymbol{t},F;\boldsymbol{x}) = \mathbb{E}[\exp(it_jF_j^T\boldsymbol{x})]$. Also define $f_X^{(j)}(F) = \frac{1}{n}\sum_i \epsilon_i\psi_j(\boldsymbol{t},F;\boldsymbol{x}^{(i)})$. Thus, using the same argument as in Lemma A.6 for $\boldsymbol{t} \leftarrow \|\boldsymbol{t}\|\boldsymbol{e}_j$ and $\boldsymbol{x}^{(i)} \leftarrow t_j\boldsymbol{x}^{(i)}$,

$$\sup_{F\in\mathcal{F}}|\psi_j(\boldsymbol{t},F|\hat{P}) - \psi_j(\boldsymbol{t},F|P)| = O_P\left(|t_j|\sqrt{\frac{k\|S\|\log n}{n}}\right)$$

Finally, we see that:

$$\begin{aligned}
\sup_{F\in\mathcal{F}}\left|\psi(\boldsymbol{t},F|\hat{P}) - \psi(\boldsymbol{t},F|P)\right| &\leq \sum_j O_P\left(|t_j|\sqrt{\frac{k\|S\|\log n}{n}}\right) \\
&= O_P\left(\sqrt{\frac{k\|S\|\log n}{n}}\right)\left(\sum_{j=1}^k|t_j|\right) \quad = \|\boldsymbol{t}\|O_P\left(\sqrt{\frac{k^2\|S\|\log n}{n}}\right)
\end{aligned}$$

The above is true because, for $|a_i|, |b_i| \leq 1$, $i = 1,\ldots,k$,

$$\left|\prod_{i=1}^k a_i - \prod_{i=1}^k b_i\right| = \left|\sum_{j=0}^{k-1}\prod_{i\leq j}b_i(a_{j+1} - b_{j+1})\prod_{i=j+2}^k a_i\right| \leq \sum_{j=0}^{k-1}|a_{j+1} - b_{j+1}|$$

□

A.1.2 Local convergence

In this section, we provide the proof of Theorem 4. Recall the ICA model from Eq 1,

$$\begin{aligned}
&\boldsymbol{x} = B\boldsymbol{z} + \boldsymbol{g}, \\
&\mathrm{diag}\,(\boldsymbol{a}) := \mathrm{cov}\,(\boldsymbol{z})\,, \quad \Sigma := \mathrm{cov}\,(\boldsymbol{g}) = B\,\mathrm{diag}\,(\boldsymbol{a})\,B^T + \Sigma
\end{aligned}$$

We provide the proof for the CGF-based contrast function. The proof for the CHF-based contrast function follows similarly. From the Proof of Theorem A.1 we have,

$$\nabla f(\boldsymbol{u}|P) = \frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right) B\boldsymbol{z}\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]} - B\,\mathrm{diag}\left(\boldsymbol{a}\right) B^T\boldsymbol{u},$$

$$\nabla^2 f(\boldsymbol{u}|P) = \frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right) B\boldsymbol{z}\boldsymbol{z}^T B^T\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]} - \frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right) B\boldsymbol{z}\right]\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right) z^T B^T\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]^2} - B\,\mathrm{diag}\left(\boldsymbol{a}\right) B^T$$

$$= B\left[\underbrace{\frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\boldsymbol{z}\boldsymbol{z}^T\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]} - \frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\boldsymbol{z}\right]\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\boldsymbol{z}^T\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]^2}}_{\text{Covariance of } \boldsymbol{z} \sim \frac{\mathbb{P}(\boldsymbol{z})\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]}} - \mathrm{diag}\left(\boldsymbol{a}\right)\right] B^T$$

The new probability density $\frac{\mathbb{P}(\boldsymbol{z})\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B\boldsymbol{z}\right)\right]}$ is an exponential tilt of the original pdf, and since $\{z_i\}_{i=1}^d$ are independent, and the new tilted density also factorizes over the z_i's, therefore, the covariance under this tilted density is also diagonal.

Let $C := BD_0B^T$. We denote the i^{th} column of B as $B_{.i}$. Define functions

$$r_i(\boldsymbol{u}) := \ln\left(\mathbb{E}\left[\exp\left(\boldsymbol{u}^T B_{.i} z_i\right)\right]\right) - \mathrm{Var}(z_i)\frac{\boldsymbol{u}^T B_{.i} B_{.i}^T \boldsymbol{u}}{2}$$
$$g_i(\boldsymbol{u}) := \nabla r_i(\boldsymbol{u})$$

Then $f(\boldsymbol{u}|P) = \sum_{i=1}^d r_i(\boldsymbol{u})$, $\nabla f(\boldsymbol{u}|P) = \sum_{i=1}^d g_i(\boldsymbol{u})$. For fixed point iteration, $\boldsymbol{u}_{k+1} = \frac{\nabla f\left(C^{-1}\boldsymbol{u}_k|P\right)}{\|\nabla f(C^{-1}\boldsymbol{u}_k|P)\|}$. Consider the function $\nabla f\left(C^{-1}\boldsymbol{u}|P\right)$. Let $\boldsymbol{e}_i$ denote the i^{th} axis-aligned unit basis vector of $\mathbb{R}^n$. Since B is full-rank, we can denote $\boldsymbol{\alpha} = B^{-1}\boldsymbol{u}$. We have,

$$\begin{aligned}
\nabla f\left(C^{-1}\boldsymbol{u}|P\right) &= \sum_{i=1}^d g_i\left(C^{-1}\boldsymbol{u}\right) \\
&= \sum_{i=1}^d \frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T\left(C^{-1}\right)^T B_{.i} z_i\right) B_{.i} z_i\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T\left(C^{-1}\right)^T B_{.i} z_i\right)\right]} - \mathrm{Var}(z_i)\, B_{.i} B_{.i}^T C^{-1}\boldsymbol{u} \\
&= \sum_{i=1}^d \frac{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T\left(B^T\right)^{-1} D_0^{-1}\boldsymbol{e}_i z_i\right) B_{.i} z_i\right]}{\mathbb{E}\left[\exp\left(\boldsymbol{u}^T\left(B^T\right)^{-1} D_0^{-1}\boldsymbol{e}_i z_i\right)\right]} - \mathrm{Var}(z_i)\, B_{.i}\boldsymbol{e}_i^T D_0^{-1} B^{-1} u \\
&= \sum_{i=1}^d \left(\frac{\mathbb{E}\left[\exp\left(\frac{\alpha_i}{(D_0)_{ii}} z_i\right) z_i\right]}{\mathbb{E}\left[\exp\left(\frac{\alpha_i}{(D_0)_{ii}} z_i\right)\right]} - \mathrm{Var}(z_i)\frac{\alpha_i}{(D_0)_{ii}}\right) B_{.i}
\end{aligned}$$

For $t \in \mathbb{R}$, define the function $q_i(t) : \mathbb{R} \to \mathbb{R}$ as -

$$q_i(t) := \frac{\mathbb{E}\left[\exp\left(\frac{t}{(D_0)_{ii}} z_i\right) z_i\right]}{\mathbb{E}\left[\exp\left(\frac{t}{(D_0)_{ii}} z_i\right)\right]} - \mathrm{Var}(z_i)\frac{t}{(D_0)_{ii}}$$

Note that $q_i(0) = \mathbb{E}[z_i] = 0$. For $\boldsymbol{t} = (t_1, t_2, \cdots t_d) \in \mathbb{R}^d$, let $\boldsymbol{q}(\boldsymbol{t}) := [q_1(t_1), q_2(t_2) \cdots q_d(t_d)]^T \in \mathbb{R}^d$. Then,

$$\nabla f\left(C^{-1}u|P\right) = B\boldsymbol{q}(\boldsymbol{\alpha})$$

 https://doi.org/10.52202/079017-4217

Therefore, if $\boldsymbol{u}_{t+1} = B\boldsymbol{\alpha}_{t+1}$, then

$$B\boldsymbol{\alpha}_{t+1} = \frac{B\boldsymbol{q}\left(\boldsymbol{\alpha}_t\right)}{\|B\boldsymbol{q}\left(\boldsymbol{\alpha}_t\right)\|}$$
$$\implies \boldsymbol{\alpha}_{t+1} = \frac{\boldsymbol{q}\left(\boldsymbol{\alpha}_t\right)}{\|B\boldsymbol{q}\left(\boldsymbol{\alpha}_t\right)\|}$$
$$\implies \forall i \in [d],\ \left(\boldsymbol{\alpha}_{t+1}\right)_i = \frac{q_i\left(\left(\boldsymbol{\alpha}_t\right)_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{\alpha}_t\right)\|}$$

At the fixed point, $\boldsymbol{\alpha}^* = \frac{\boldsymbol{e}_1}{\|B\boldsymbol{e}_1\|}$ and $\frac{B\boldsymbol{q}\left(\boldsymbol{\alpha}^*\right)}{\|B\boldsymbol{q}\left(\boldsymbol{\alpha}^*\right)\|_2} = B\boldsymbol{e}_1$. Therefore,

$$\forall i \in [2,d], \left(\boldsymbol{\alpha}_{t+1}\right)_i - \alpha_i^* = \left(\boldsymbol{\alpha}_{t+1}\right)_i = \frac{q_i\left(\left(\boldsymbol{\alpha}_t\right)_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{\alpha}_t\right)\|} \tag{A.26}$$

$$\text{for } i = 1, \left(\boldsymbol{\alpha}_{t+1}\right)_1 - \alpha_1^* = \frac{q_1\left(\left(\boldsymbol{\alpha}_t\right)_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{\alpha}_t\right)\|} - \frac{1}{\|B\boldsymbol{e}_1\|} \tag{A.27}$$

Note the smoothness assumptions on $q_i(.)$ mentioned in Theorem 4, $\forall i \in [d]$,

1. $\sup_{t\in[-\|B^{-1}\|_2,\|B^{-1}\|_2]}|q_i\left(t\right)| \leq c_1$
2. $\sup_{t\in[-\|B^{-1}\|_2,\|B^{-1}\|_2]}|q_i'\left(t\right)| \leq c_2$
3. $\sup_{t\in[-\|B^{-1}\|_2,\|B^{-1}\|_2]}|q_i''\left(t\right)| \leq c_3$

Since $\forall k,\ \|\boldsymbol{u}_k\| = 1$, therefore, $\forall k,\ \|\boldsymbol{\alpha}_t\| \leq \|B^{-1}\|_2$. We seek to prove that the sequence $\{\boldsymbol{\alpha}_t\}_{k=1}^n$ converges to $\boldsymbol{\alpha}^*$.

A.1.2.1 Taylor expansions

Consider the function $g_i\left(\boldsymbol{y}\right) := \frac{q_i\left(y_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}$ for $\boldsymbol{y} \in \mathbb{R}^d$. We start by computing the gradient, $\nabla_{\boldsymbol{y}} g_i\left(\boldsymbol{y}\right)$.

Lemma A.8. *Let* $g_i\left(\boldsymbol{y}\right) := \frac{q_i\left(\boldsymbol{y}_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}$*, then we have*

$$\left[\nabla_{\boldsymbol{y}} g_i\left(\boldsymbol{y}\right)\right]_j = \begin{cases} \frac{1}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}\frac{\partial q_i\left(y_i\right)}{\partial y_i} - \frac{q_i\left(y_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|^2}\frac{\partial\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}{\partial q_i\left(y_i\right)}\frac{\partial q_i\left(y_i\right)}{\partial y_i}, & \textit{for } j = i \\ -\frac{q_j'\left(y_j\right)}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}\frac{q_i\left(y_i\right)\boldsymbol{e}_j^T B^T B\boldsymbol{q}\left(\boldsymbol{y}\right)}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|^2}, & \textit{for } j \neq i \end{cases}$$

Proof. The derivative w.r.t y_i is given as -

$$\begin{aligned}\frac{\partial g_i\left(\boldsymbol{y}\right)}{\partial y_i} &= \frac{1}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}\frac{\partial q_i\left(y_i\right)}{\partial y_i} - \frac{q_i\left(y_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|^2}\frac{\partial\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}{\partial y_i} \\ &= \frac{1}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}\frac{\partial q_i\left(y_i\right)}{\partial y_i} - \frac{q_i\left(y_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|^2}\frac{\partial\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}{\partial q_i\left(y_i\right)}\frac{\partial q_i\left(y_i\right)}{\partial y_i}\end{aligned}$$

Note that

$$\nabla_{\boldsymbol{y}}\|A\boldsymbol{y}\|_2 = \frac{1}{\|A\boldsymbol{y}\|}A^T A\boldsymbol{y}$$

Therefore,

$$\begin{aligned}\frac{\partial g_i\left(\boldsymbol{y}\right)}{\partial y_i} &= \frac{1}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}\frac{\partial q_i\left(y_i\right)}{\partial y_i} - \frac{q_i\left(y_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|^2}\frac{1}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}\boldsymbol{e}_i^T B^T B\boldsymbol{q}\left(\boldsymbol{y}\right)\frac{\partial q_i\left(y_i\right)}{\partial y_i} \\ &= \frac{q_i'\left(y_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|}\left[1 - \frac{q_i\left(y_i\right)\boldsymbol{e}_i^T B^T B\boldsymbol{q}\left(\boldsymbol{y}\right)}{\|B\boldsymbol{q}\left(\boldsymbol{y}\right)\|^2}\right]\end{aligned} \tag{A.28}$$

For $j \neq i$, the derivative w.r.t y_j is given as -

$$\begin{aligned}\frac{\partial g_i(\boldsymbol{y})}{\partial y_j} &= -\frac{q_i(y_i)}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2}\frac{\partial\|B\boldsymbol{q}(\boldsymbol{y})\|}{\partial q_j(y_j)}\frac{\partial q_j(y_j)}{\partial y_j} \\ &= -\frac{q_j'(y_j)}{\|B\boldsymbol{q}(\boldsymbol{y})\|}\frac{q_i(y_i)\,\boldsymbol{e}_j^T B^T B\boldsymbol{q}(\boldsymbol{y})}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2} \end{aligned} \tag{A.29}$$

□

Next, we bound $q_i(t)$ and $q_i'(t)$.

Lemma A.9. *Under the smoothness assumptions on $q_i(.)$ mentioned in Theorem 4, we have for $t \in [-\|B^{-1}\|_2, \|B^{-1}\|_2]$,*

1. $|q_1(t) - q_1(\alpha_1^*)| \leq c_2|t-\alpha_1^*|, \;\; |q_1'(t) - q_1'(\alpha_1^*)| \leq c_3|t-\alpha_1^*|$
2. $\forall i,\, |q_i(t)| \leq \frac{c_3 t^2}{2}, \;\; |q_i'(t)| \leq c_3|t|$

Proof. First consider $q_i(t)$ and $q_i'(t)$ for $i \neq 1$. Using Taylor expansion around $t=0$, we have for some $c \in (0,t)$ and $\forall i \neq 1$,

$$\begin{aligned} q_i(t) &= q_i(0) + tq_i'(0) + \frac{t^2}{2}q_i''(c) \text{ and} \\ q_i'(t) &= q_i'(0) + tq_i''(0) \end{aligned}$$

Now, we know that $q_i(0) = q_i'(0) = 0$. Then using Assumption 1, we have, for $t \in [-\|B^{-1}\|_2, \|B^{-1}\|_2]$,

$$|q_i(t)| \leq \frac{c_3 t^2}{2} \text{ and} \tag{A.30}$$

$$|q_i'(t)| \leq c_3|t| \tag{A.31}$$

Similarly, using Taylor expansion around $\alpha_1^* = \frac{1}{\|B\boldsymbol{e}_1\|_2}$ for $q_1(.)$ we have, for some $c' \in (0,t)$

$$\begin{aligned} q_1(t) &= q_1(\alpha_1^*) + (t-\alpha_1^*)\,q_1'(c') \text{ and} \\ q_1'(t) &= q_1'(\alpha_1^*) + (t-\alpha_1^*)\,q_1''(c') \end{aligned}$$

Therefore, using Assumption 4, we have, for $t \in [-\|B^{-1}\|_2, \|B^{-1}\|_2]$

$$|q_1(t) - q_1(\alpha_1^*)| \leq c_2|t-\alpha_1^*| \text{ and} \tag{A.32}$$

$$|q_1'(t) - q_1'(\alpha_1^*)| \leq c_3|t-\alpha_1^*| \tag{A.33}$$

□

A.1.2.2 Using convergence radius

In this section, we use the Taylor expansion results for $q_i(.)$ and $q_i'(.)$ in the previous section to analyze the following functions for $\boldsymbol{y} \in \mathbb{R}^d$,

1. $w(\boldsymbol{y}) := \|B\boldsymbol{y}\|$
2. $v(\boldsymbol{y}; i) := \dfrac{\boldsymbol{e}_i^T B^T B\boldsymbol{q}(\boldsymbol{y})}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2}$

Under the constraints specified in the theorem statement we have,

$$\begin{aligned} &\|\boldsymbol{y} - \boldsymbol{\alpha}^*\|_2 \leq R, \;\; R \leq \max\{c_2/c_3, 1\}, \\ &\epsilon := \frac{\|B\|_F}{\|B\boldsymbol{e}_1\|}\max\left\{\frac{c_2}{|q_1(\alpha_1^*)|}, \frac{c_3}{|q_1'(\alpha_1^*)|}\right\}, \;\; \epsilon R \leq \frac{1}{10} \end{aligned} \tag{A.34}$$

We start with $w(\boldsymbol{y})$.

 https://doi.org/10.52202/079017-4217

Lemma A.10. $\forall\, \boldsymbol{y} \in \mathbb{R}^d$, *satisfying* (A.34)*, let* $\delta(\boldsymbol{y}) := \boldsymbol{q}(\boldsymbol{y}) - q_1(\alpha_1^*)\boldsymbol{e}_1$*. Then, we have*

1. $(1-\epsilon\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|)\,|q_1(\alpha_1^*)|\,\|B\boldsymbol{e}_1\| \le \|B\boldsymbol{q}(\boldsymbol{y})\| \le (1+\epsilon\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|)\,|q_1(\alpha_1^*)|\,\|B\boldsymbol{e}_1\|$
2. $\|\delta(\boldsymbol{y})\| \le c_2\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|$

Proof. Consider the vector $\delta(\boldsymbol{y}) := \boldsymbol{q}(\boldsymbol{y}) - q_1(\alpha_1^*)\boldsymbol{e}_1$. Then,

$$|(\delta(\boldsymbol{y}))_\ell| \le \begin{cases} c_2\,|y_1-\alpha_1^*|, \text{ for } \ell=1, \text{ using Eq A.32} \\ \frac{c_3}{2}(y)_\ell^2, \text{ for } \ell\neq 1, \text{ using Eq A.30} \end{cases} \tag{A.35}$$

Note that

$$\|\delta(\boldsymbol{y})\| \le c_2\|\boldsymbol{y}-\boldsymbol{\alpha}^*\| \tag{A.36}$$

Now consider $w(\boldsymbol{y})$. Using the mean-value theorem on the Euclidean norm we have,

$$\|B\boldsymbol{q}(\boldsymbol{y})\| = |q_1(\alpha_1^*)|\,\|B\boldsymbol{e}_1\| + \frac{1}{\|B\gamma(\boldsymbol{y})\|}\left(B^TB\gamma(\boldsymbol{y})\right)^T\delta(\boldsymbol{y}), \tag{A.37}$$

where $\gamma(\boldsymbol{y}) = \mu\boldsymbol{q}(\boldsymbol{y}) + (1-\mu)\,q_1(\alpha_1^*)\boldsymbol{e}_1,\ \mu\in(0,1)$. Then,

$$\begin{aligned}
\left\|\frac{1}{\|B\gamma(\boldsymbol{y})\|}\left(B^TB\gamma(\boldsymbol{y})\right)^T\delta(\boldsymbol{y})\right\| &\le \frac{1}{\|B\gamma(\boldsymbol{y})\|}\left\|B^TB\gamma(\boldsymbol{y})\right\|\|\delta(\boldsymbol{y})\| \\
&\le \frac{1}{\|B\gamma(\boldsymbol{y})\|}\|B\|\,\|B\gamma(\boldsymbol{y})\|\,\|\delta(\boldsymbol{y})\| \\
&= \|B\|\,\|\delta(\boldsymbol{y})\| \\
&\le c_2\|B\|\,\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|, \text{ using Eq A.36} \\
&\le c_2\|B\|_F\,\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|, \\
&\le \epsilon\,|q_1(\alpha_1^*)|\,\|B\boldsymbol{e}_1\|\,\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|, \text{ using Eq A.34}
\end{aligned} \tag{A.38}$$

Therefore using A.37,

$$(1-\epsilon\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|)\,|q_1(\alpha_1^*)|\,\|B\boldsymbol{e}_1\| \le \|B\boldsymbol{q}(\boldsymbol{y})\| \le (1+\epsilon\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|)\,|q_1(\alpha_1^*)|\,\|B\boldsymbol{e}_1\| \tag{A.39}$$

□

Finally, we consider the function $v(\boldsymbol{y}; i) := \dfrac{\boldsymbol{e}_i^TB^TB\boldsymbol{q}(\boldsymbol{y})}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2}$.

Lemma A.11. $\forall\, \boldsymbol{y}\in\mathbb{R}^d$ *satisfying* (A.34)*, we have*

1. *For* $i=1,\ 1-5\epsilon\|\boldsymbol{y}-\boldsymbol{\alpha}^*\| \le q_1(\alpha_1^*)\,v(\boldsymbol{y}|1) \le 1+5\epsilon\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|$
2. *For* $i\neq 1,\ |q_1(\alpha_1^*)\,v(\boldsymbol{y};i)| \le (1+5\epsilon\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|)\dfrac{\|B\boldsymbol{e}_i\|}{\|B\boldsymbol{e}_1\|}$

Proof. Define $\theta := \epsilon\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|$ for convenience of notation. We have,

$$\frac{\boldsymbol{e}_i^TB^TB\boldsymbol{q}(\boldsymbol{y})}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2} = q_1(\alpha_1^*)\frac{\boldsymbol{e}_i^TB^TB\boldsymbol{e}_1}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2} + \frac{\boldsymbol{e}_i^TB^TB\delta(\boldsymbol{y})}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2}, \text{ using definition of } \delta(\boldsymbol{y}) \tag{A.40}$$

Therefore, if $i=1$, then using Lemma A.10,

$$\frac{1}{(1+\theta)^2} + \frac{q_1(\alpha_1^*)\,\boldsymbol{e}_1^TB^TB\delta(\boldsymbol{y})}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2} \le \frac{q_1(\alpha_1^*)\,\boldsymbol{e}_1^TB^TB\boldsymbol{q}(\boldsymbol{y})}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2} \le \frac{1}{(1-\theta)^2} + \frac{q_1(\alpha_1^*)\,\boldsymbol{e}_1^TB^TB\delta(\boldsymbol{y})}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2} \tag{A.41}$$

Using the following inequalities -

$$\frac{1}{(1+\theta)^2} \ge 1-2\theta, \theta\ge 0, \text{ and, } \frac{1}{(1-\theta)^2} \le 1+5\theta, \theta \le \frac{1}{5}$$

and observing that using Eq A.34, and Lemma A.10,

$$\begin{aligned}\left|\frac{q_1(\alpha_1^*)\boldsymbol{e}_1^T B^T B\delta(\boldsymbol{y})}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2}\right| &\leq |q_1(\alpha_1^*)| \frac{\|B\boldsymbol{e}_1\|}{(1-\theta)^2 |q_1(\alpha_1^*)|^2 \|B\boldsymbol{e}_1\|^2} \|B\| \|\delta(\boldsymbol{y})\| \\ &\leq \frac{1}{(1-\theta)^2} \frac{c_2\|B\|}{|q_1(\alpha_1^*)| \|B\boldsymbol{e}_1\|} \|\boldsymbol{y}-\boldsymbol{\alpha}^*\| \\ &\leq \frac{\epsilon}{(1-\theta)^2} \|\boldsymbol{y}-\boldsymbol{\alpha}^*\|\end{aligned}$$

Therefore we have from Eq A.41,

$$1-5\theta \leq q_1(\alpha_1^*) v(\boldsymbol{y};i) \leq 1+5\theta \tag{A.42}$$

where we used $\theta := \epsilon\|\boldsymbol{y}-\boldsymbol{\alpha}^*\|$. For the case of $i \neq 1$, using Lemma A.10 we have

$$\begin{aligned}|q_1(\alpha_1^*)||v(\boldsymbol{y};i)| &= |q_1(\alpha_1^*)| \left|\frac{\boldsymbol{e}_i^T B^T B\boldsymbol{q}(\boldsymbol{y})}{\|B\boldsymbol{q}(\boldsymbol{y})\|^2}\right| \\ &\leq |q_1(\alpha_1^*)| \left|\frac{\|B\boldsymbol{e}_i\|}{\|B\boldsymbol{q}(\boldsymbol{y})\|}\right| \\ &\leq \frac{1}{(1-\theta)^2} \frac{\|B\boldsymbol{e}_i\|}{\|B\boldsymbol{e}_1\|} \\ &\leq (1+5\theta) \frac{\|B\boldsymbol{e}_i\|}{\|B\boldsymbol{e}_1\|}\end{aligned} \tag{A.43}$$

□

We now operate under the assumption that Eq A.34 holds for $\boldsymbol{y} = \boldsymbol{\alpha}_t$ and inductively show that it holds for $\boldsymbol{y} = \boldsymbol{\alpha}_{t+1}$ as well. Recall the function $g_i(\boldsymbol{y}) := \frac{q_i(y_i)}{\|B\boldsymbol{q}(\boldsymbol{y})\|}$. By applying the mean-value theorem for $g_i(.)$ for the points $\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}^*$, we have from Eq A.26 and A.27 -

$$\begin{aligned}|(\boldsymbol{\alpha}_{t+1})_i - \alpha_i^*| &= |g_i(\boldsymbol{\alpha}_t) - g_i(\boldsymbol{\alpha}^*)| \\ &= \left|\nabla g_i(\boldsymbol{\beta}_i)^T (\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*)\right| \text{ for } \boldsymbol{\beta}_i := (1-\lambda_i)\boldsymbol{\alpha}_t + \lambda_i\boldsymbol{\alpha}^*,\ \lambda_i \in (0,1) \\ &\leq \|\nabla g_i(\boldsymbol{\beta}_i)\| \|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*\|\end{aligned} \tag{A.44}$$

Note the induction hypothesis assumes that Eq A.34 is true for $x = \boldsymbol{\alpha}_t$. Since $\forall i, \lambda_i \in (0,1)$, therefore Eq A.34 holds for all $\boldsymbol{\beta}_i$ as well. Squaring and adding Eq A.44 for $i \in [d]$ and taking a square-root, we have

$$\begin{aligned}\|\boldsymbol{\alpha}_{t+1} - \boldsymbol{\alpha}^*\| &\leq \|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*\| \sqrt{\sum_{i=1}^k \|\nabla g_i(\boldsymbol{\beta}_i)\|^2} \\ &\leq \|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*\| \sqrt{\sum_{i=1}^k \sum_{j=1}^k \left(\frac{\partial g_i(\boldsymbol{y})}{\partial y_j}\bigg|_{\boldsymbol{\beta}_i}\right)^2}\end{aligned} \tag{A.45}$$

Let us consider the expression $G_{ij} := \frac{\partial g_i(\boldsymbol{y})}{\partial y_j}\bigg|_{\boldsymbol{\beta}_i}, \forall i,j \in [k]$. For the purpose of the subsequent analysis, we define $\theta := \epsilon\|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*\|$. We divide the analysis into the following cases -

Case 1 : $i = 1, j = 1$

Using Lemma A.8,

$$|G_{11}| = \left|\frac{q_1'((\boldsymbol{\beta}_1)_1)}{\|Bw(\boldsymbol{\beta}_1)\|}\left[1 - \frac{q_1((\boldsymbol{\beta}_1)_1)\boldsymbol{e}_1^T B^T B\boldsymbol{q}(\boldsymbol{\beta}_1)}{\|B\boldsymbol{q}(\boldsymbol{\beta}_1)\|^2}\right]\right|$$

https://doi.org/10.52202/079017-4217

From Lemma A.9,

$$\begin{aligned}
|q_1'((\boldsymbol{\beta}_1)_1)| &\leq |q_1'(\alpha_1^*)| + c_3 |(\boldsymbol{\beta}_1)_1 - \alpha_1^*| \\
&= |q_1'(\alpha_1^*)| \left(1 + c_3 \frac{|(\boldsymbol{\beta}_1)_1 - \alpha_1^*|}{|q_1'(\alpha_1^*)|}\right) \\
&\leq |q_1'(\alpha_1^*)| \left(1 + c_3 \frac{\|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*\|}{|q_1'(\alpha_1^*)|}\right) \\
&\leq |q_1'(\alpha_1^*)| (1+\theta)
\end{aligned}$$

and,

$$\begin{aligned}
q_1((\boldsymbol{\beta}_1)_1) &\leq |q_1(\alpha_1^*)| + c_2 |(\boldsymbol{\beta}_1)_1 - \alpha_1^*| \\
&= |q_1(\alpha_1^*)| \left(1 + c_2 \frac{|(\boldsymbol{\beta}_1)_1 - \alpha_1^*|}{|q_1(\alpha_1^*)}\right) \\
&\leq |q_1(\alpha_1^*)| \left(1 + c_2 \frac{\|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*\|}{|q_1(\alpha_1^*)|}\right) \\
&\leq |q_1(\alpha_1^*)| (1+\theta)
\end{aligned}$$

From Lemma A.10,

$$\|B\boldsymbol{q}(\boldsymbol{\beta}_1)\| \geq (1-\theta) |q_1(\alpha_1^*)| \|B\boldsymbol{e}_1\|$$

From Lemma A.11 and $\theta \leq \frac{1}{10}$,

$$-6\theta \leq 1 - \frac{q_1((\boldsymbol{\beta}_1)_1) \boldsymbol{e}_1^T B^T B w(\boldsymbol{\beta}_1)}{\|Bw(\boldsymbol{\beta}_1)\|^2} \leq 6\theta$$

Therefore,

$$\begin{aligned}
|G_{11}| &\leq 6 \left(\frac{1+\theta}{1-\theta}\right) \frac{|q_1'(\alpha_1^*)| \theta}{\|B\boldsymbol{e}_1\| |q_1(\alpha_1^*)|} \\
&\leq \frac{7.5\epsilon |q_1'(\alpha_1^*)|}{\|B\boldsymbol{e}_1\| |q_1(\alpha_1^*)|} \|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*\|, \text{ since } \theta \leq \frac{1}{10}
\end{aligned}$$

Case 2 : $i = 1, j \neq 1$

Using Lemma A.8,

$$|G_{1j}| = \left| \frac{q_j'\left((\boldsymbol{\beta}_1)_j\right)}{\|B\boldsymbol{q}(\boldsymbol{\beta}_1)\|} \frac{q_1((\boldsymbol{\beta}_1)_1) \boldsymbol{e}_j^T B^T B \boldsymbol{q}(\boldsymbol{\beta}_1)}{\|B\boldsymbol{q}(\boldsymbol{\beta}_1)\|^2} \right|$$

From Lemma A.9,

$$\left|q_j'\left((\boldsymbol{\beta}_1)_j\right)\right| \leq c_3 \left|(\boldsymbol{\beta}_1)_j - \alpha_j^*\right| \leq c_3 \left|(\boldsymbol{\alpha}_t)_j - \alpha_j^*\right|, \text{ since } \alpha_j^* = 0$$

From Lemma A.10,

$$\|B\boldsymbol{q}(\boldsymbol{\beta}_1)\| \geq (1-\theta) |q_1(\alpha_1^*)| \|B\boldsymbol{e}_1\|$$

From Lemma A.9,

$$\begin{aligned}
|q_1((\boldsymbol{\beta}_1)_1)| &\leq |q_1(\alpha_1^*)| + c_2 |(\boldsymbol{\beta}_1)_1 - \alpha_1^*| \\
&= |q_1(\alpha_1^*)| \left(1 + c_2 \frac{|(\boldsymbol{\beta}_1)_1 - \alpha_1^*|}{|q_1(\alpha_1^*)|}\right) \\
&\leq |q_1(\alpha_1^*)| (1+\theta)
\end{aligned}$$

From Lemma A.11,

$$\left| \frac{\boldsymbol{e}_j^T B^T B \boldsymbol{q}(\boldsymbol{\beta}_1)}{\|B\boldsymbol{q}(\boldsymbol{\beta}_1)\|^2} \right| \leq \frac{(1+5\theta)}{|q_1(\alpha_1^*)|} \frac{\|B\boldsymbol{e}_j\|}{\|B\boldsymbol{e}_1\|}$$

Therefore,

$$|G_{1j}| \leq c_3 \left(\frac{1+\theta}{1-\theta}\right)(1+5\theta)\frac{\|B\boldsymbol{e}_j\|}{|q_1(\alpha_1^*)|\|B\boldsymbol{e}_1\|^2}\left|(\boldsymbol{\alpha}_t)_j - \alpha_j^*\right| \leq 2c_3\frac{\|B\boldsymbol{e}_j\|}{|q_1(\alpha_1^*)|\|B\boldsymbol{e}_1\|^2}\left|(\boldsymbol{\alpha}_t)_j - \alpha_j^*\right|$$

Case 3 : $i \neq 1, j = 1$

Using Lemma A.8,

$$|G_{i1}| = \left|\frac{q_1'((\boldsymbol{\beta}_i)_1)}{\|B\boldsymbol{q}(\boldsymbol{\beta}_i)\|}\frac{q_i((\boldsymbol{\beta}_i)_i)\boldsymbol{e}_1^T B^T B\boldsymbol{q}(\boldsymbol{\beta}_i)}{\|B\boldsymbol{q}(\boldsymbol{\beta}_i)\|^2}\right|$$

From Lemma A.9,

$$\begin{aligned}|q_1'((\boldsymbol{\beta}_i)_1)| &\leq |q_1'(\alpha_1^*)| + c_3|(\boldsymbol{\beta}_i)_1 - \alpha_1^*| \\ &= |q_1'(\alpha_1^*)|\left(1 + c_3\frac{|(\boldsymbol{\beta}_i)_1 - \alpha_1^*|}{|q_1'(\alpha_1^*)|}\right) \\ &\leq |q_1'(\alpha_1^*)|(1+\theta)\end{aligned}$$

From Lemma A.10,

$$\|B\boldsymbol{q}(\boldsymbol{\beta}_i)\| \geq (1-\theta)|q_1(\alpha_1^*)|\|B\boldsymbol{e}_1\|$$

From Lemma A.9,

$$|q_i((\boldsymbol{\beta}_i)_i)| \leq \frac{c_3}{2}((\boldsymbol{\beta}_i)_i - \alpha_i^*)^2 \leq \frac{c_3}{2}((\boldsymbol{\alpha}_t)_i - \alpha_i^*)^2$$

From Lemma A.11,

$$\left|\frac{\boldsymbol{e}_1^T B^T B\boldsymbol{q}(\boldsymbol{\beta}_i)}{\|B\boldsymbol{q}(\boldsymbol{\beta}_i)\|^2}\right| \leq \frac{1+5\theta}{|q_1(\alpha_1^*)|}$$

Therefore,

$$\begin{aligned}|G_{i1}| &\leq \frac{c_3}{2}\left(\frac{(1+\theta)(1+5\theta)}{1-\theta}\right)\frac{|q_1'(\alpha_1^*)|}{|q_1(\alpha_1^*)|^2\|B\boldsymbol{e}_1\|}((\boldsymbol{\alpha}_t)_i - \alpha_i^*)^2 \\ &\leq c_3\frac{|q_1'(\alpha_1^*)|}{|q_1(\alpha_1^*)|^2\|B\boldsymbol{e}_1\|}((\boldsymbol{\alpha}_t)_i - \alpha_i^*)^2 \\ &\leq c_3 R\frac{|q_1'(\alpha_1^*)|}{|q_1(\alpha_1^*)|^2\|B\boldsymbol{e}_1\|}|(\boldsymbol{\alpha}_t)_i - \alpha_i^*| \\ &\leq c_2\frac{|q_1'(\alpha_1^*)|}{|q_1(\alpha_1^*)|^2\|B\boldsymbol{e}_1\|}|(\boldsymbol{\alpha}_t)_i - \alpha_i^*|\end{aligned}$$

Case 4 : $i \neq 1, j \neq 1, i = j$

Using Lemma A.8,

$$|G_{ii}| = \left|\frac{q_i'((\boldsymbol{\beta}_i)_i)}{\|B\boldsymbol{q}(\boldsymbol{\beta}_i)\|}\left[1 - \frac{q_i((\boldsymbol{\beta}_i)_i)\boldsymbol{e}_i^T B^T B\boldsymbol{q}(\boldsymbol{\beta}_i)}{\|B\boldsymbol{q}(\boldsymbol{\beta}_i)\|^2}\right]\right|$$

From Lemma A.9,

$$|q_i'((\boldsymbol{\beta}_i)_i)| \leq c_3\left|(\boldsymbol{\beta}_i)_i - \alpha_j^*\right| \leq c_3|(\boldsymbol{\alpha}_t)_i - \alpha_i^*|$$

From Lemma A.10,

$$\|B\boldsymbol{q}(\boldsymbol{\beta}_i)\| \geq (1-\theta)|q_1(\alpha_1^*)|\|B\boldsymbol{e}_1\|$$

From Lemma A.9,

$$|q_i((\boldsymbol{\beta}_i)_i)| \leq \frac{c_3}{2}((\boldsymbol{\beta}_i)_i - \alpha_i^*)^2 \leq \frac{c_3}{2}((\boldsymbol{\alpha}_t)_i - \alpha_i^*)^2 \leq c_3 R\|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*\| \leq c_2\|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}^*\|$$

 https://doi.org/10.52202/079017-4217

From Lemma A.11,

$$\left|\frac{\boldsymbol{e}_i^T B^T B\boldsymbol{q}\left(\boldsymbol{\beta}_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{\beta}_i\right)\|^2}\right| \leq (1+5\theta)\frac{\|B\boldsymbol{e}_i\|}{|q_1\left(\alpha_1^*\right)|\|B\boldsymbol{e}_1\|}$$

Then,

$$\begin{aligned}\frac{q_i\left(\left(\boldsymbol{\beta}_i\right)_i\right)\boldsymbol{e}_i^T B^T B\boldsymbol{q}\left(\boldsymbol{\beta}_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{\beta}_i\right)\|^2} &\leq (1+5\theta)\left\|\boldsymbol{\alpha}_t-\boldsymbol{\alpha}^*\right\|\frac{\|B\boldsymbol{e}_i\|}{\|B\boldsymbol{e}_1\|}\frac{c_2}{|q_1\left(\alpha_1^*\right)|}\\ &\leq (1+5\theta)\left\|\boldsymbol{\alpha}_t-\boldsymbol{\alpha}^*\right\|\frac{\|B\|_F}{\|B\boldsymbol{e}_1\|}\frac{c_2}{|q_1\left(\alpha_1^*\right)|}\\ &\leq (1+5\theta)\,\epsilon\left\|\boldsymbol{\alpha}_t-\boldsymbol{\alpha}^*\right\|\\ &\leq 2\theta\end{aligned}$$

Therefore,

$$|G_{ii}| \leq \frac{c_3\left|\left(\boldsymbol{\alpha}_t\right)_i-\alpha_i^*\right|}{(1-\theta)\left|q_1\left(\alpha_1^*\right)\right|\|B\boldsymbol{e}_1\|} \leq \frac{2c_3}{\left|q_1\left(\alpha_1^*\right)\right|\|B\boldsymbol{e}_1\|}\left|\left(\boldsymbol{\alpha}_t\right)_i-\alpha_i^*\right|$$

Case 5 : $i\neq 1, j\neq 1, i\neq j$

Using Lemma A.8,

$$|G_{ij}| = \left|\frac{q_j'\left(\left(\boldsymbol{\beta}_i\right)_j\right)}{\|B\boldsymbol{q}\left(\left(\boldsymbol{\beta}_i\right)\right)\|}\frac{q_i\left(\left(\boldsymbol{\beta}_i\right)_i\right)\boldsymbol{e}_j^T B^T B\boldsymbol{q}\left(\left(\boldsymbol{\beta}_i\right)\right)}{\|B\boldsymbol{q}\left(\left(\boldsymbol{\beta}_i\right)\right)\|^2}\right|$$

From Lemma A.9,

$$\begin{aligned}\left|q_j'\left(\left(\boldsymbol{\beta}_i\right)_j\right)\right| &\leq c_3\left|\left(\boldsymbol{\beta}_i\right)_j-\alpha_j^*\right|\\ &\leq c_3\left|\left(\boldsymbol{\alpha}_t\right)_j-\alpha_j^*\right|\end{aligned}$$

From Lemma A.10,

$$\|B\boldsymbol{q}\left(\boldsymbol{\beta}_i\right)\| \geq (1-\theta)\left|q_1\left(\alpha_1^*\right)\right|\|B\boldsymbol{e}_1\|$$

From Lemma A.9,

$$\left|q_i\left(\left(\boldsymbol{\beta}_i\right)_i\right)\right| \leq \frac{c_3}{2}\left(\left(\boldsymbol{\beta}_i\right)_i-\alpha_i^*\right)^2 \leq \frac{c_3}{2}\left(\left(\boldsymbol{\alpha}_t\right)_i-\alpha_i^*\right)^2 \leq \frac{c_3 R}{2}\left|\left(\boldsymbol{\alpha}_t\right)_i-\alpha_i^*\right|$$

From Lemma A.11,

$$\left|\frac{\boldsymbol{e}_j^T B^T B\boldsymbol{q}\left(\boldsymbol{\beta}_i\right)}{\|B\boldsymbol{q}\left(\boldsymbol{\beta}_i\right)\|^2}\right| \leq \frac{(1+5\theta)}{\left|q_1\left(\alpha_1^*\right)\right|}\frac{\|B\boldsymbol{e}_j\|}{\|B\boldsymbol{e}_1\|}$$

Therefore,

$$|G_{ij}| \leq \frac{c_3^2 R}{2}\left(\frac{1+5\theta}{1-\theta}\right)\frac{\|B\boldsymbol{e}_j\|}{\left|q_1\left(\alpha_1^*\right)\right|^2\|B\boldsymbol{e}_1\|^2}\left|\left(\boldsymbol{\alpha}_t\right)_j-\alpha_j^*\right|\left|\left(\boldsymbol{\alpha}_t\right)_i-\alpha_i^*\right| \leq \frac{c_3^2 R\|B\boldsymbol{e}_j\|}{\left|q_1\left(\alpha_1^*\right)\right|^2\|B\boldsymbol{e}_1\|^2}\left|\left(\boldsymbol{\alpha}_t\right)_j-\alpha_j^*\right|\left|\left(\boldsymbol{\alpha}_t\right)_i-\alpha_i^*\right|$$

A.1.2.3 Putting everything together

Putting all the cases together in Eq A.45, we have $\frac{\|\boldsymbol{\alpha}_{t+1}-\boldsymbol{\alpha}^*\|}{\|\boldsymbol{\alpha}_t-\boldsymbol{\alpha}^*\|} \leq C\left\|\boldsymbol{\alpha}_t-\boldsymbol{\alpha}^*\right\|$, where

$$C := \sqrt{\begin{aligned}&\left(\frac{7.5\epsilon\left|q_1'\left(\alpha_1^*\right)\right|}{\|B\boldsymbol{e}_1\|\left|q_1\left(\alpha_1^*\right)\right|}\right)^2+\left(\frac{2c_3\|B\|_F}{\left|q_1\left(\alpha_1^*\right)\right|\|B\boldsymbol{e}_1\|^2}\right)^2+\left(\frac{c_2\left|q_1'\left(\alpha_1^*\right)\right|}{\left|q_1\left(\alpha_1^*\right)\right|^2\|B\boldsymbol{e}_1\|}\right)^2+\left(\frac{2c_3}{\left|q_1\left(\alpha_1^*\right)\right|\|B\boldsymbol{e}_1\|}\right)^2\\&+\left(\frac{c_3^2R^2\|B\|_F}{\left|q_1\left(\alpha_1^*\right)\right|^2\|B\boldsymbol{e}_1\|^2}\right)^2\end{aligned}}$$

Then we have, $C \leq \frac{5\|B\|_F}{\|B\boldsymbol{e}_1\|^2}\max\left\{\frac{c_3}{\left|q_1\left(\alpha_1^*\right)\right|}, \frac{c_2^2}{\left|q_1\left(\alpha_1^*\right)\right|^2}\right\}$. For linear convergence, we require $CR<1$, which is ensured by the condition mentioned in Theorem 4.

A.2 Additional experiments for noisy ICA

A.2.1 Experiments with super-gaussian source signals

Table 2 shows the Amari error in the presence of super-Gaussian signals. The signals are 1) a Uniform distribution $U(-\sqrt{3}, \sqrt{3})$, 2) Bernoulli$\left(\frac{1}{2} + \frac{1}{\sqrt{12}}\right)$, 3) Laplace(0, 0.05) (mean 0 and standard deviation 0.05), 4) Exponential(5), and 5) and 6) a Student's t-distribution with 3 and 5 degrees of freedom, respectively. The Meta algorithm closely follows the best candidate algorithm even in the presence of many super-Gaussian signals.

Table 2: Variation of Amari error with Sample Size for Heavy-Tailed distributions, averaged over 100 random runs. Noise power $\rho = 0.001$, number of sources $k = 6$.

Algorithm \ n	200	500	1000	5000	10000
Meta	**0.44376**	**0.25215**	**0.20222**	**0.11435**	**0.0838**
CHF	1.10103	0.70823	0.52529	0.21344	**0.09194**
CGF	1.84266	1.51216	1.3702	0.84351	0.58753
PEGI	2.27873	1.86561	1.71474	1.33709	1.23322
PFICA	**0.39237**	**0.25468**	**0.21222**	**0.12878**	0.12347
JADE	0.70174	0.39246	0.28199	0.12652	0.09686
FastICA	0.66441	0.3869	0.28419	0.13215	0.09946

A.2.2 Image demixing experiments

In this section, we provide additional experiments for image-demixing using ICA. We mix images using flattening and linearly mixing them with a 4×4 matrix B (i.i.d entries $\sim \mathcal{N}(0, 1)$) and Wishart noise ($\rho = 0.001$). Demixing is performed using the SINR-optimal demixing matrix (see Section 2) and the results are shown in Figure A.1 along with their corresponding independence scores. The CHF-based method recovers the original sources well, upto sign. The Kurtosis and CGF-based method fails to recover the second source. This is consistent with their higher independence score. The Meta algorithm selects CHF from candidates CHF, CGF, Kurtosis, FastICA, and JADE.

A.2.3 Image denoising experiments

In this experiment, we use the ICA-based denoising technique proposed in [38] to compare candidate Noisy ICA algorithms and show that the Meta algorithm can pick the best-denoised image based on the independence score proposed in our work.

We use the noisy MNIST dataset and further add entrywise Gaussian noise with variance proportional to $\sin^2(c_1\pi(i+j))(c_1 = 50)$ to the $(i,j)^{\text{th}}$ pixel. Training images are flattened to create a 784-dimensional vector, and PCA is performed to reduce dimensionality to 25, on which subsequently ICA is performed. The original and denoised images along with their independence score are shown in Figure A.2. We note that CHF-based denoising provides qualitatively better results, while CGF-based denoising provides the worst results. This is consistent with their corresponding independence scores.

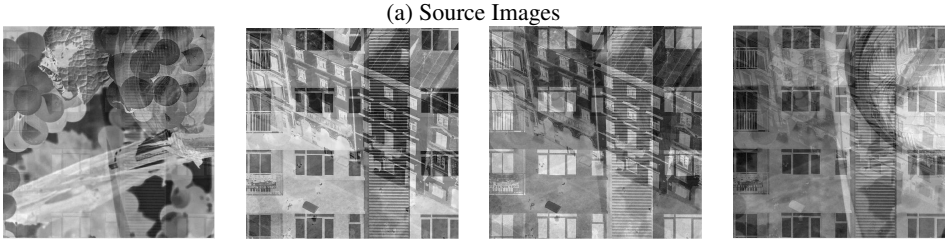
A.3 Algorithm for sequential calculation of independence scores

In this section, we provide an algorithm for the computation of the independence score proposed in the manuscript, but in a sequential manner. The detailed algorithm is provided as Algorithm 3. We assume that we have access to a matrix of the form $C = BDB^T$.

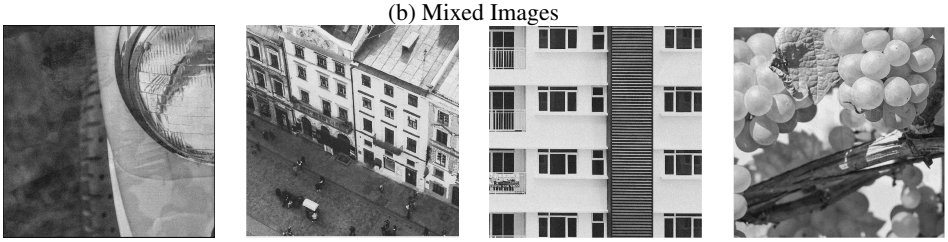
The power method (see Eq 3) essentially extracts one column of B (up to scaling) at every step. [49] provides an elegant way to use the pseudo-Euclidean space to successively project the data onto columns orthogonal to the ones extracted. This enables one to extract each column of the mixing matrix. For completeness, we present the steps of this projection. After extracting the i^{th}

https://doi.org/10.52202/079017-4217

(a) Source Images



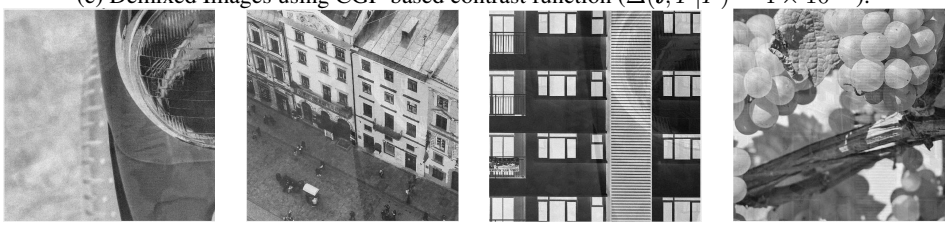
(b) Mixed Images



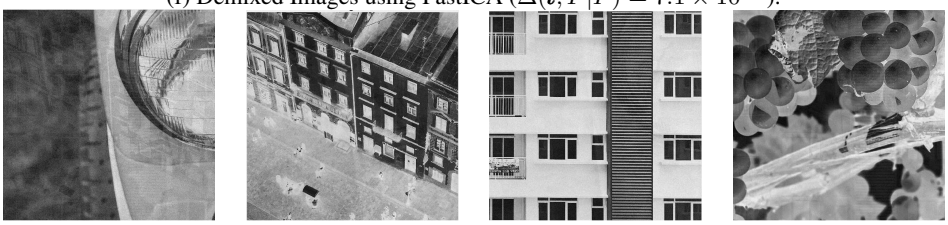
(c) Demixed Images using CHF-based contrast function ($\Delta(\boldsymbol{t}, F|P) = 5.26 \times 10^{-3}$).

(d) Demixed Images using Kurtosis-based contrast function ($\Delta(\boldsymbol{t}, F|P) = 2.48 \times 10^{-2}$)

(e) Demixed Images using CGF-based contrast function ($\Delta(\boldsymbol{t}, F|P) = 4 \times 10^{-2}$).



(f) Demixed Images using FastICA ($\Delta(\boldsymbol{t}, F|P) = 7.1 \times 10^{-3}$).



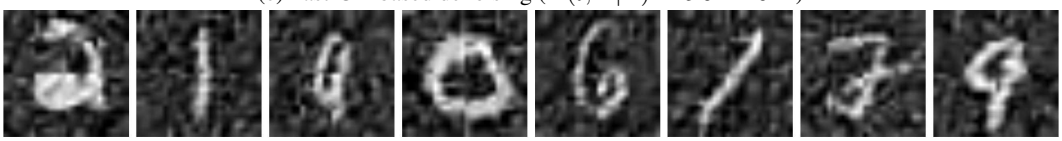
(a) Original Images

(b) CHF-based denoising ($\Delta(\boldsymbol{t}, F|P) = 4.4 \times 10^{-6}$)

(c) Kurtosis-based denoising ($\Delta(\boldsymbol{t}, F|P) = 1.7 \times 10^{-4}$)

(d) CGF-based denoising ($\Delta(\boldsymbol{t}, F|P) = 9.2 \times 10^{-6}$)

(e) FastICA-based denoising ($\Delta(\boldsymbol{t}, F|P) = 5.9 \times 10^{-6}$)



(f) JADE-based denoising ($\Delta(\boldsymbol{t}, F|P) = 4.6 \times 10^{-6}$)

Figure A.2: Image Denoising using ICA

column $\boldsymbol{u}^{(i)}$, the algorithm maintains two matrices. The first, denoted by U, estimates the mixing matrix B one column at a time (up to scaling). The second, denoted by V estimates B^{-1} *one row* at a time. It is possible to extend the independence score in Eq 2 to the sequential setting as follows. Let $\boldsymbol{u}^{(j)}$ be the j^{th} vector extracted by the power method (Eq 3) until convergence. After extracting ℓ columns, we project the data using $\min(\ell + 1, k)$ projection matrices. These would be mutually independent if we indeed extracted different columns of B. Let $C = BDB^T$ for some diagonal matrix D. For convenience of notation, denote $B_i \equiv B(:, i)$. Set -

$$\boldsymbol{v}^{(j)} = \frac{C^\dagger \boldsymbol{u}^{(j)}}{{\boldsymbol{u}^{(j)}}^T C^\dagger \boldsymbol{u}^{(j)}} \tag{A.46}$$

When $\boldsymbol{u}^{(j)} = B_j/\|B_j\|$, $\boldsymbol{v}^{(j)} = \dfrac{(B^T)^{-1} D^\dagger \boldsymbol{e}_j}{\boldsymbol{e}_j^T D^\dagger \boldsymbol{e}_j} \|B_j\| = (B^T)^{-1} \boldsymbol{e}_j \|B_j\|$. Let $\boldsymbol{x}$ denote an arbitrary datapoint. Thus

$$U(:, j)V(j, :)\boldsymbol{x} = B_j z_j + U(:, j)V(j, :)\boldsymbol{g} \tag{A.47}$$

Thus the projection on all other columns $j > \ell$ is given by:

$$(I - UV)\boldsymbol{x} = \sum_{i=1}^{k} B_i z_i - \sum_{j=1}^{\ell} B_j z_j + \tilde{\boldsymbol{g}} = \sum_{j=\ell+1}^{k} B_j z_j + \tilde{\boldsymbol{g}}$$

https://doi.org/10.52202/079017-4217

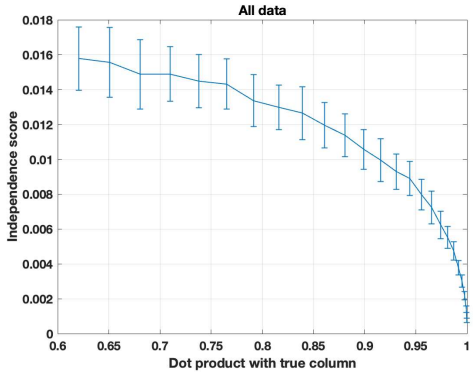


Figure A.3: Mean independence score with errorbars from 50 random runs

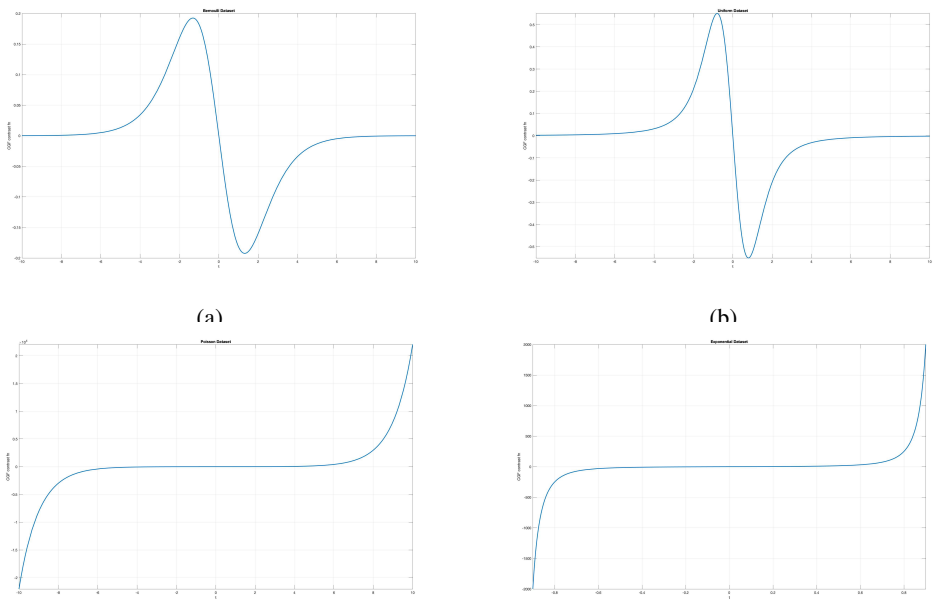
where $\tilde{\boldsymbol{g}} = F\boldsymbol{g}$, where F is some $k \times k$ matrix. So we have $\min(\ell, k)$ vectors of the form $z_i B_i + \tilde{\boldsymbol{g}}_i$, and when $\ell < k$ an additional vector which contains all independent random variables z_j, $j > \ell$ along with a mean-zero Gaussian vector. Then we can project each vector to a scalar using unit random vectors and check for independence using the Independence Score defined in 2. When $\ell = k$, then all we need are the $j = 1, \ldots, k$ projections on the k directions identified via ICA in Eq A.47.

As an example, we conduct an experiment (Figure A.3), where we fix a mixing matrix B using the same generating mechanism in Section 5. Now we create vectors $\mathbf{q}$ which interpolate between $B(:, 1)$ and a fixed arbitrary vector orthogonal to $B(:, 1)$. As the interpolation changes, we plot the independence score for direction $\mathbf{q}$. The dataset is the 9-dimensional dataset, which has independent components from many different distributions (see Section 5).

This plot clearly shows that there is a clear negative correlation between the score and the dot product of a vector with the column $B(:, 1)$. To be concrete, when $\mathbf{q}$ has a small angle with $B(:, 1)$, the independence score is small, and the error bars are also very small. However, as the dot product decreases, the score grows and the error bars become larger.

Algorithm 3 Independence Score after extracting ℓ columns. U and V are running estimates of B and B^{-1} upto ℓ columns and rows respectively.

Input
$X \in \mathbb{R}^{n \times k}$, $U \in \mathbb{R}^{k \times \ell}$, $V \in \mathbb{R}^{\ell \times k}$, Number of random projections M
$k_0 \leftarrow \min(\ell + 1, k)$
for j in range[1, ℓ] **do**
  $Y_j \leftarrow X \left(U(:, j)V(j, :)\right)^T$
end for
if $\ell < k$ **then**
  $Y_{\ell+1} \leftarrow X \left(I - V^T U^T\right)$
end if
for j in range[1, M] **do**
  $\boldsymbol{t} \leftarrow$ random unit vector in $\mathbb{R}^k$
  for a in range[1, k_0] **do**
    $W(:, j) \leftarrow Y_j \boldsymbol{t}$
  end for
  Let $\boldsymbol{\alpha} \in \mathbb{R}^{1, k_0}$ represent a row of W
  $\tilde{S} \leftarrow \text{cov}(W)$
  $\gamma \leftarrow \sum_{i,j} \tilde{S}_{ij}$
  $\boldsymbol{\beta} \leftarrow \sum_i \tilde{S}_{ii}$
  $s(j) = |\hat{\mathbb{E}} \exp(\sum_j i\alpha_j) \exp(-\boldsymbol{\beta}) - \prod_{j=1}^{k_0} \hat{\mathbb{E}} \exp(i\alpha_j) \exp(-\gamma)|$
end for
Return mean(s), stdev(s)



(a) (b)

(c) (d)

Figure A.4: Plots of the third derivative of the CGF-based contrast function for different datasets - Bernoulli($p = 0.5$) A.4a, Uniform($\mathcal{U}\left(-\sqrt{3}, \sqrt{3}\right)$) A.4b, Poisson($\lambda = 1$) A.4c and Exponential($\lambda = 1$) A.4d. Note that the sign stays the same in each half-line

We refer to this function as $\Delta\left(X, U, V, \ell\right)$ which takes as input, the data X, the number of columns ℓ and matrices U, V which are running estimates of B and B^{-1} upto ℓ columns and rows respectively.

A.4 More details for third derivative condition in theorem 3

Theorem 3 requires the condition "The third derivative of $h_X(u)$ does not change the sign in the half line $[0, \infty)$ for the non-Gaussian random variable considered in the ICA problem.". In this section, we provide sufficient conditions with respect to contrast functions and datasets where this holds. Further, we also provide an interesting example to demonstrate that our Assumption 1-(d) might not be too far from being necessary.

CGF-based contrast function. Consider the cumulant generating function of a random variable X, i.e. the logarithm of the moment generating function. Now consider the contrast function

$$g_X(t) = CGF(t) - \text{var}(X)t^2/2.$$

We first note that it satisfies Assumption 1(a)-(d). Next, we observe that it is enough for a distribution to have all cumulants of the same sign to satisfy Assumption 1(d) for the CGF. For example, the Poisson distribution has all positive cumulants. Figure A.4 depicts the third derivative of the contrast function for a Bernoulli(1/2), Uniform, Poisson, and Exponential.

Logarithm of the symmetrized characteristic function. Figure A.5 depicts the third derivative of this contrast function for the logistic distribution where Assumption 1(d) holds. Although the Assumption does not hold for a large class of distributions here, we believe that the notion of having the same sign can be relaxed up to some bounded parameter values instead of the entire half line, so that the global convergence results of Theorem 3 still hold. This belief is further reinforced by Figure A.6 which shows that the loss landscape of functions where Assumption 1(d) doesn't hold is still smooth.

https://doi.org/10.52202/079017-4217

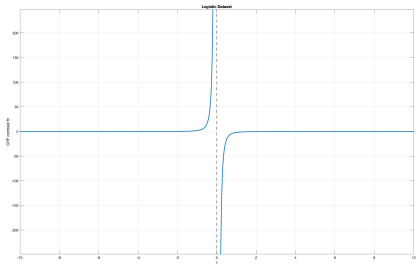


Figure A.5: Plots of the third derivative of the CHF-based contrast function for Logistic$(0, 1)$ dataset.

A.4.1 Towards a necessary condition

Consider the probability distribution function (see [36]) $f(x) = ke^{-\frac{x^2}{2}}(a+b\cos(c\pi x))$ for constants $a, b, c > 0$. For $f(x)$ to be a valid pdf, we require that $f(x) \geq 0$, and $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Therefore, we have,

$$\int_{-\infty}^{\infty} ke^{-\frac{x^2}{2}}(a + b\cos(c\pi x))\,dx = 1 \tag{A.48}$$

Noting standard integration results, we have that,

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx = \sqrt{2\pi} \quad \text{and} \quad \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\cos(c\pi x)\,dx = \sqrt{2\pi}e^{-\frac{c^2\pi^2}{2}} \tag{A.49}$$

Therefore, $k = \frac{1}{\sqrt{2\pi}(a+be^{-\frac{c^2\pi^2}{2}})}$. Some algebraic manipulations yield the following:

$$MGF_{f(x)}(t) = \frac{1}{(a + be^{-\frac{c^2\pi^2}{2}})}e^{\frac{t^2}{2}}\left(a + be^{-\frac{c^2\pi^2}{2}}\cos(c\pi t)\right)$$

Therefore, $\ln(MGF_{f(x)}(t)) = -\ln(a + be^{-\frac{c^2\pi^2}{2}}) + \frac{t^2}{2} + \ln(a + be^{-\frac{c^2\pi^2}{2}}\cos(c\pi t))$.

We now compute the mean, μ, and variance σ^2.

The mean, μ can be given as :

$$\mu = \int_{-\infty}^{\infty} f(x)x\,dx = 0 \text{ since } f(x) \text{ is an even function.}$$

The variance, σ^2 can therefore be given as :

$$\sigma^2 = k\left(a\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}x^2\,dx + b\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}x^2\cos(c\pi x)\,dx\right)$$

Now, we note that $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}x^2\,dx = \sqrt{2\pi}$ and $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}x^2\cos(c\pi x)\,dx = e^{-\frac{c^2\pi^2}{2}}\sqrt{2\pi}(1-c^2\pi^2)$
Therefore,

$$\sigma^2 = k\left(a\sqrt{2\pi} + be^{-\frac{c^2\pi^2}{2}}\sqrt{2\pi}(1-c^2\pi^2)\right) = 1 - \frac{bc^2\pi^2e^{-\frac{c^2\pi^2}{2}}}{a + be^{-\frac{c^2\pi^2}{2}}}$$

The symmetrized CGF can therefore be written as

$$\begin{aligned}\operatorname{sym} CGF(t) &:= \ln(MGF_{f(x)}(t)) + \ln(MGF_{f(x)}(-t)) \\ &= -2\ln(a + be^{-\frac{c^2\pi^2}{2}}) + t^2 + 2\ln(a + be^{-\frac{c^2\pi^2}{2}}\cos(c\pi t))\end{aligned}$$

Therefore, $g(t) = \operatorname{sym} CGF(t) - (1 - \frac{bc^2\pi^2e^{-\frac{c^2\pi^2}{2}}}{a+be^{-\frac{c^2\pi^2}{2}}})t^2$ can be written as :

$$g(t) = -2\ln(a + be^{-\frac{c^2\pi^2}{2}}) + 2\ln(a + be^{-\frac{c^2\pi^2}{2}}\cos(c\pi t)) + \frac{bc^2\pi^2e^{-\frac{c^2\pi^2}{2}}}{a + be^{-\frac{c^2\pi^2}{2}}}t^2$$

Finally, $g'(t)$ can be written as :

$$g'(t) = \frac{d}{dt}\left(2\ln(a + be^{-\frac{c^2\pi^2}{2}}\cos(c\pi t)) + \frac{bc^2\pi^2 e^{-\frac{c^2\pi^2}{2}}}{a + be^{-\frac{c^2\pi^2}{2}}}t^2\right)$$
$$= -\frac{2bc\pi e^{-\frac{c^2\pi^2}{2}}\sin(c\pi t)}{a + be^{-\frac{c^2\pi^2}{2}}\cos(c\pi t)} + \frac{2bc^2\pi^2 e^{-\frac{c^2\pi^2}{2}}}{a + be^{-\frac{c^2\pi^2}{2}}}t$$

$g''(t)$ can be written as :

$$g''(t) = -2bc^2\pi^2 e^{-\frac{c^2\pi^2}{2}}\frac{a\cos(c\pi t) + be^{-\frac{c^2\pi^2}{2}}}{(a + be^{-\frac{c^2\pi^2}{2}}\cos(c\pi t))^2} + \frac{2bc^2\pi^2 e^{-\frac{c^2\pi^2}{2}}}{a + be^{-\frac{c^2\pi^2}{2}}}$$

and $g'''(t)$ can be evaluated as:

$$g'''(t) = 2bc^3\pi^3 e^{-\frac{c^2\pi^2}{2}}\sin(c\pi t)\frac{\left(a^2 - 2b^2e^{-c^2\pi^2} - abe^{-\frac{c^2\pi^2}{2}}\cos(c\pi t)\right)}{(a + be^{-\frac{c^2\pi^2}{2}}\cos(c\pi t))^3}$$

Lets set $a = 2, b = -1, c = \frac{4}{\pi}$. Then, we have that the pdf, $f(x) = ke^{-\frac{x^2}{2}}(2 - \cos(4x)) = ke^{-\frac{x^2}{2}}(1 + 2\sin^2(2x))$. The corresponding functions are :

$$E[\exp(tX)] = \frac{1}{(2 - e^{-8})}e^{\frac{t^2}{2}}\left(2 - e^{-8}\cos(4t)\right)$$

$$g(t) := \operatorname{sym} CGF(t) - \operatorname{var}(X)t^2 = -2\ln(2 - e^{-8}) + 2\ln(2 - e^{-8}\cos(4t)) - \frac{16e^{-8}}{2 - e^{-8}}t^2$$

$$g'(t) = \frac{8e^{-8}\sin(4t)}{2 - e^{-8}\cos(4t)} - \frac{32e^{-8}}{2 - e^{-8}}t$$

$$g''(t) = 32e^{-8}\frac{2\cos(4t) - e^{-8}}{(2 - e^{-8}\cos(4t))^2} - \frac{32e^{-8}}{2 - e^{-8}}$$

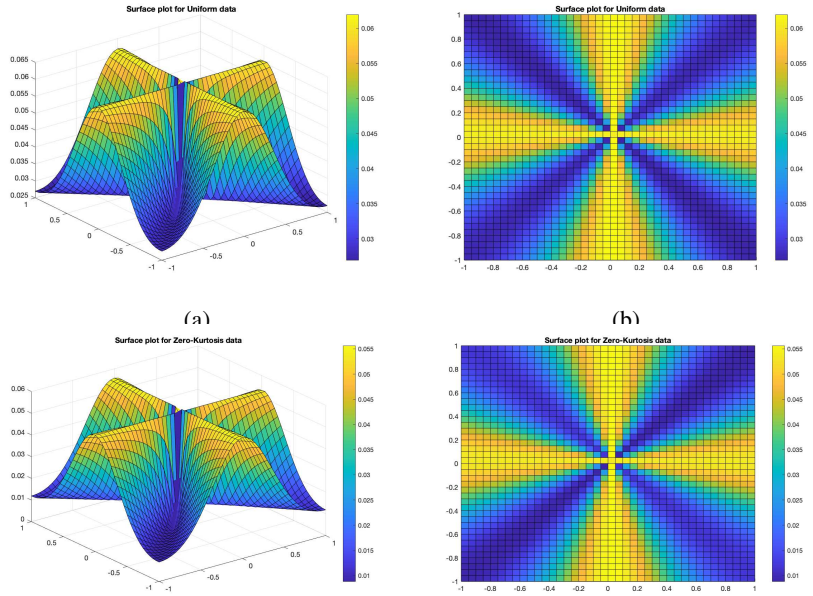
$$g'''(t) = -128e^{-8}\sin(4t)\frac{\left(4 - 2e^{-16} + 2e^{-8}\cos(4t)\right)}{(2 - e^{-8}\cos(4t))^3}$$

Closeness to a Gaussian MGF: It can be seen there that $g'''(t)$ changes sign in the half-line. However, the key is to note that the MGF of this distribution is very close to that of a Gaussian and can be made arbitrarily small by varying the parameters a and b. This renders the CGF-based contrast function ineffectual. This leads us to believe that Assumption 1(d) is not far from being necessary to ensure global optimality of the corresponding objective function.

A.5 Surface plots of the CHF-based contrast functions

Figure A.6 depicts the loss landscape of the Characteristic function (CHF) based contrast function described in Section 3.3 of the manuscript. We plot the value of the contrast function evaluated at $B^{-1}\boldsymbol{u}, \boldsymbol{u} = \frac{1}{\sqrt{x^2+y^2}}\begin{pmatrix}x\\y\end{pmatrix}$ for $x, y \in [-1, 1]$. As shown in the figure. We have rotated the data so that the columns of B align with the X and Y axes. The global maxima occur at $\boldsymbol{u}$ aligned with the columns of B.

https://doi.org/10.52202/079017-4217



(a) (b)

(c) (d)

Figure A.6: Surface plots for zero-kurtosis (A.6c and A.6d) and Uniform (A.6a, A.6b) data with $n = 10000$ points, noise-power $\rho = 0.1$ and number of source signals, $k = 2$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main contributions are the Meta algorithm for non-parametric selection of the best noisy ICA algorithm for a given distribution, along with 2 new contrast functions of independent interest.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the Experiments (Section 5, there are some noise/sample-size regimes where other algorithms may perform slightly better than our proposed contrast functions, but we show that even then our diagnostic can pick out the best algorithm.

   Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Every theorem statement mentions the assumptions and the location of the proof in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 provides extensive details of the experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

 https://doi.org/10.52202/079017-4217

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We make the code for our experiments available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 5 provides extensive details of the experimental setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Figure 1 and 2(a) show error bars and error distribution respectively. Figure 2(b),(c) show the mean over 100 runs, similar to [49]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Our experiments were performed on a single Macbook Pro M2 2022 CPU with 8 GB RAM.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

 https://doi.org/10.52202/079017-4217

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We abide by the NeurIPS Code of Ethics in our work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any models with a risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: MATLAB implementations (under the GNU General Public License) can be found at - FastICA and JADE. The code for PFICA was provided on request by the authors.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets to require documentation or licensing.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We only use publicly available datasets in Section 5.

https://doi.org/10.52202/079017-4217

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not have experiments with crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.